GaN HEMT Modeling and Design for Millimeter and Sub-millimeter Wave Power Amplifiers through Monte Carlo Particle-based Device Simulations

by

Diego Guerra

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2011 by the
Graduate Supervisory Committee:

Marco Saraniti, Chair
David K. Ferry
Stephen M. Goodnick
Sule Ozev

ARIZONA STATE UNIVERSITY

December 2011

ABSTRACT

The drive towards device scaling and large output power in millimeter and sub-millimeter wave power amplifiers results in a highly non-linear, out-of-equilibrium charge transport regime. Particle-based Full Band Monte Carlo device simulators allow an accurate description of this carrier dynamics at the nanoscale.

This work initially compares *GaN* high electron mobility transistors (HEMTs) based on the established *Ga*-face technology and the emerging *N*-face technology, through a modeling approach that allows a fair comparison, indicating that the *N*-face devices exhibit improved performance with respect to *Ga*-face ones due to the natural back-barrier confinement that mitigates short-channel-effects. An investigation is then carried out on the minimum aspect ratio (*i.e.* gate length to gate-to-channel-distance ratio) that limits short channel effects in ultra-scaled *GaN* and *InP* HEMTs, indicating that this value in *GaN* devices is 15 while in *InP* devices is 7.5. This difference is believed to be related to the different dielectric properties of the two materials, and the corresponding different electric field distributions. The dielectric effects of the passivation layer in millimeter-wave, high-power *GaN* HEMTs are also investigated, finding that the effective gate length is increased by fringing capacitances, enhanced by the dielectrics in regions adjacent to the gate for layers thicker than 5 *nm*, strongly affecting the frequency performance of deep sub-micron devices.

Lastly, efficient Full Band Monte Carlo particle-based device simulations of the large-signal performance of mm-wave transistor power amplifiers with high-Q matching networks are reported for the first time. In particular, a Cellular Monte Carlo (CMC) code is self-consistently coupled with a Harmonic Balance (HB) frequency domain circuit solver. Due to the iterative nature of the HB algorithm, this simulation approach is possible only due to the computational efficiency of the CMC, which uses pre-computed scattering tables. On the other hand, HB allows the direct simulation of

the steady-state behavior of circuits with long transient time. This work provides an accurate and efficient tool for the device early-stage design, which allows a computer-based performance evaluation in lieu of the extremely time-consuming and expensive iterations of prototyping and experimental large-signal characterization.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Marco Saraniti for his continual support, and for constantly providing me with motivations and challenges. I am also very grateful for his trust and mentoring in the pursuit of my interests and ideas. I thank Dr. Stephen M. Goodnick for his guidance and support, Dr. David K. Ferry for the his enlightening observations during the research group meetings, Dr. Richard Akis, and Dr Sule Ozev. I would like to acknowledge the funding that supported this work provided by the Air Force Research Laboratory (AFRL) National High Reliability Electronics Virtual Center (HiREV) under Wyle Laboratory Contract managed by Stephen Tetlak.

I would also like to thank my friends and colleagues Dr. Fabio Alessio Marino, Dr. Nicolas Faralli, Riccardo Soligo, and Flavio Maria Sabatti, for sharing this doctoral adventure with me, and for their stimulating and entertaining discussions during those long days spent in the lab. Finally, I would particularly like to thank April, and my family and friends back in Italy for their support and understanding throughout the past three years.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

INTRODUCTION 1

1.1 The Cellular Monte Carlo Technique 3

1.2 Purpose of the Work 8

2. HIGH-FREQUENCY CHARACTERIZATION OF MILLIMETER AND SUBMILLIMETER-BAND FET AMPLIFIERS 10

2.1 Introduction 10

2.2 FET Two-port Network Model. 10

2.2.1 Physical Meaning of f_T 14

2.3 Small-signal FET Equivalent AC Circuit 16

2.4 FET RF Small-signal Amplifiers 20

2.5 FET RF Power Amplifiers 25

2.5.1 Non-linear Effects 32

2.5.2 Large-signal versus Small-signal Analysis 35

2.5.3 Power Amplifier Figures of Merit 38

2.5.4 RF Power Amplifier Classes 41

3. EMERGING N-FACE GAN HEMT TECHNOLOGY 52

3.1 Introduction 52

3.2 Polarization Engineering in GaN HEMTs 54

3.3 Device Structure and Simulation Method 60

3.4 Gate Length Scaling 63

3.4.1 Gate Length Scaling Effect on DC Performance 63

3.4.2 Gate Length Scaling Effect on RF Performance 65

3.5 Impact of Access Region on Device Performance 65

3.6 Enhancement and Depletion mode 68

3.7 Conclusions 69

4. COMPARISON OF N-FACE AND GA-FACE GAN HEMTS 70

4.1 Introduction 70

4.2 Comparison between N-face and Ga-face Technology 70

4.3 Back Barrier Effect on Device Scaling 76

4.3.1 Gate Length Scaling Effect on DC Performance 76

4.3.2 Gate Length Scaling Effect on RF Performance 78

4.4 Conclusions 79

5. HIGHLY-SCALED HEMT OPTIMIZATION FOR IMPROVED DC AND HIGH-FREQUENCY PERFORMANCE 80

5.1 Introduction 80

5.2 Effects of Aspect Ratio Scaling on DC performance 81

5.3 Effects of Aspect Ratio Scaling on RF performance 84

5.4 Role of Aspect Ratio and Effective Gate Length 86

5.5 Access Region Scaling for Improved Performance 93

5.6 Conclusions 95

CHAPTER Page

6. EFFECTS OF THE PASSIVATION DIELECTRIC CONSTANT ON THE RF PERFORMANCE OF MILLIMITER-WAVE POWER GAN HEMTS 96

6.1 Introduction 96

6.2 Simulation Approach 98

6.3 Effective Gate Length and Cut-off frequency 100

6.4 Carrier Dynamics along the Channel 103

6.5 High-power Biasing 110

6.6 Feedback Capacitance and Output Impedance 116

6.7 Conclusions 121

7. CIRCUIT-DEVICE SIMULATIONS 122

7.1 Introduction 122

7.2 AC Small-signal Network Solver 123

7.3 Finite-Difference Time-Domain (FDTD) Network Solver 127

7.3.1 FDTD Network Topologies 129

7.3.2 Real-time Smoothing Algorithm 136

7.3.3 Down-sampling 138

7.4 Harmonic Balance Frequency-Domain Network Solver 140

7.4.1 The Key Idea of the Harmonic Balance algorithm 141

7.4.2 The General Harmonic Balance Algorithm 145

7.4.3 The Self-consistent Harmonic Balance / Monte Carlo Algorithm 148

7.4.4 Windowing Effect and Single Iteration Length 151

CHAPTER Page

7.4.5 Convergence Improvement under Relaxation 152

7.4.6 Difference between the HB Solver and the other Network Solvers 154

7.4.7 HB input modes 156

7.5 Conclusion 158

8. LARGE-SIGNAL CHARACTERIZATION OF MMWAVE GAN HEMT POWER AMPLIFIERS 159

8.1 Introduction 159

8.2 Time-efficient Simulation of mm-wave Power Amplifiers with High-Q Output Matching Networks: The Active Load-line Technique 160

8.2.1 The High-Q Input Matching Network 166

8.3 Inclusion of Parasitic Elements in Large-signal operations 168

8.4 Large-signal Simulations of mm-wave GaN HEMT Power Amplifiers 171

8.5 Characterization of a State-of-the-art mm-wave InAlN/*GaN* HEMT Power Amplifier 180

8.6 Conclusion 189

9. CONCLUSIONS 190

REFERENCES 192

LIST OF TABLES

Table Page

3.1 Material 3.1 Ideal and predicted real c/a ratio comparison for binary wurtzite group III nitrides [1] 55

LIST OF FIGURES

Figure Page

1.1 Material properties important for the device high-frequency, high-power performance [2]. The parameter JM is the Johnson's figure of merit [3]. 2

1.2 Monte Carlo simulation: a) Illustration of the simulation flow. b) Example of possible final states to be chosen during the scattering pro- -cess. c) Particle tracking in the momentum space. 6

1.3 Quantum mechanical confinement effects (charge setback and energy level quantization) in a triangular potential well obtained through effe- -ctive potential approach. 7

2.1 Equivalent two-port network of the FET based on the Y-parameters. 11

2.2 Full-frequency Y-parameter extraction through step perturbations app- -lied through two separate simulations. 12

2.3 Estimation of f_T and f_{max} from, respectively, the 0 dB intercept point of the short-circuit current gain $h_{21}(f)$ and the maximum unilateral trans- -ducer gain, $G_{TUmax}(f)$. 13

2.4 Schematic example of effective gate length measurement based on carrier velocity (solid line) and density (dashed line) profiles along the channel. 15

2.5 FET small-signal equivalent circuit including extrinsic parasitic compo- -nents. 17

2.6 Simplified intrinsic FET small-signal equivalent circuit. 18

Figure Page

2.7 Estimation of C_{gd} from the the imaginary part of Y_{12} obtained by step perturbation. The spreading of the different curves is due to the diffe- -rent windows used for the Fourier transform. 19

2.8 Microwave amplifier block diagram. 20

2.9 Typical small-signal amplifier constant gain, G, contour plots for diffe- -rent load reflection coefficients in the Smith chart. 24

2.10 Example of a dynamic load-line overlaid to the static I_D–V_D. 26

2.11 Output impedance measurement set-up and resulting drain current sw- -ing versus drain voltage swing overlaid to the I_D–V_D. 27

2.12 Example of different load-lines whose extremes are bound to the high- -est and lowest current set by the gate voltage swing. 29

2.13 Load matching, R_L, for Maximum Small-signal Gain ($R_L = R_{out}$) versus Maximum Large-signal Output Power ($R_L = R_{opt}$) 31

2.14 Output spectrum of a non-linear system, stimulated by a two-tones input, showing the higher harmonics and the intermodulation product spurious frequency components up to the $3rd$ order. 34

2.15 Example of superposition of input tones in a linear (left) and non-linear system (right). 35

2.16 Example of constant output power contours, P_{out} , plot in the Smith chart for different input power levels, P_{in}. 38

2.17 Example of typical single-tone input large signal figures of merit: Output power (P_{out}), gain, and power added efficiency (PAE) vs input power, P_{in} at the fundamental frequency. 39

Figure Page

2.18 Example of typical two-tone input large signal figures of merit: Fundamental frequency output power and third-order intermodulation frequency output power vs input power, P_{in}. 41

2.19 Typical Class-A RF power amplifier output matching configuration. . . . 42

2.20 Example of Class-A output current and voltage waveforms (left) and dynamic load-line overlaid to the static $I_D - V_D$ (right). 43

2.21 Example of experimental large-signal characterization of a 100-mm-wide *GaN* HEMT biased near Class-A operations [4].. 43

2.22 Typical Class-B RF power amplifier output matching configuration. 44

2.23 Example of Class-B output current and voltage waveforms at the device output contacts (left) and large-signal characterization (right) 46

2.24 LEFT: Example of a typical push-pull amplifiers topology for deep Class-AB operations [5]. RIGHT:Example of experimental large-signal power performance 46

2.25 Typical Class-F RF power amplifier output matching configuration with the parallel resonant tanks tuned at the 1st, 3rd, and 5th harmonic. . . 48

2.26 Example of Class-F output current waveform and output voltage waveform. 49

2.27 Maximum efficiency of Class-F Power Amplifiers [7, 8]. 50

3.1 Comparison between typical Ga-face and N-face HEMT vertical struct--ures. 53

3.2 Schematic drawing of the crystal structure of wurtzite Ga-face and N-face *GaN* [9]. 59

Figure Page

3.3 Polarization induced sheet charge density and directions of the spontaneous and piezoelectric polarization in Ga- and N-face strained and relaxed *AlGaN*/*GaN* heterostructures [9]. 60

3.4 LEFT: Schematic cross-section of the simulated *N*-face HEMT. RIGHT: Conduction band profile, and carrier distribution 61

3.5 LEFT: Comparison of simulated (lines) and measured (dots) I_D−V_D characteristics of the HEMT. RIGHT:I_D-V_G characteristic and the corre- -sponding transconductance 62

3.6 I_D −V_D characteristics (solid lines), and transconductance (dashed lines) of the *GaN* HEMT analyzed obtained varying the gate length from L_G = 700 *nm* to L_G = 100 *nm*. 64

3.7 Electric field along the channel obtained by CMC simulations performed for V_D = 10 *V* and V_G = 0 *V*, with gate length from L_G = 700 *nm* to L_G =50 *nm*. 64

3.8 Cut-off frequencies of the N-face *GaN* HEMT extracted by the CMC simulation performed for V_D = 10 *V* and a gate biased corresponding to the peak transconductance 66

3.9 Electron drift velocity along the channel obtained by CMC simulations performed for V_D = 10 *V* and V_G corresponding to the peak transcon- -ductance 66

3.10 I_D−V_G characteristics (solid line) and respective transconductance of the *GaN* HEMT. 67

Figure Page

3.11 Electric field along the channel (L_G = 700nm) obtained by CMC simulations performed for V_D = 10 V and V_G = 0 V, scaling by an equal factor the lengths of both access regions. 67

3.12 LEFT: Conduction band profile for the Depletion (solid line) and Enhancement mode (dashed line) configurations. RIGHT: I_D-V_G (solid line) and g_m-V_G (dashed line) characteristic 69

4.1 Comparison of the simulated N-face (left) and Ga-face (right) HEMT layouts 71

4.2 Conduction band profile, and carrier distribution (black points) corresponding to a bias of 0 V on the gate and 0 V on the drain of both N-face (solid line) and Ga-face (dashed line) devices. 72

4.3 LEFT: Comparison of simulated and experimental [10] I_D−V_D characteristics of the N-face HEMT. RIGHT: Simulated I_D-V_G characte- -ristic and corresponding transconductance 73

4.4 Charge density along the direction perpendicular to the channel for the three cases described in the text, for a bias of V_G =V_D = 0 V. 74

4.5 Simulated I_D-V_G characteristic and corresponding transconductance for V_D =10 V (upper panel) and V_D = 5 V (lower panel). The N-face device and the three different Ga-face device configurations described in the text are compared. 75

4.6 Simulated g_m-V_G curves for V_D = 10 V for different device gate length (L_G = 700 nm to L_G =150 nm). 77

Figure Page

4.7 Simulated g_m-V_G (left) and I_D-V_G (right) curves for V_D = 5 V for different device gate length (L_G = 700 nm to L_G = 150 nm). The N-face device (solid line) and the Ga-face device configuration with the same gate-to-2DEG capacitance (dashed line) are compared. 77

4.8 Cut-off frequencies of the N- (solid line) and Ga-face GaN HEMTs, with the same gate-to-2DEG capacitance. 78

5.1 $InGaAs/InAlAs$ HEMT with $InGaAs$ quantum well channel simulated layout. 81

5.2 Comparison of DC short-channel effects for the GaN (V_D = 10 V) and $InGaAs$ (V_D = 2 V) devices. 82

5.3 GaN HEMT cut-off frequency for d = 15 nm versus 1/Lg (solid line) and versus 1/L_{eff} (dashed line). V_g corresponds to the peak gm, and V_d = 10 V. 85

5.4 $InGaAs$ HEMTs with fixed d = 15 nm (triangles) and fixed L_g/d = 5 (gradients) cut-off frequency versus 1/L_g. V_g corresponds to the peak g_m, and V_d = 2 V. 85

5.5 $InGaAs$ HEMTs with fixed d = 15 nm (triangles) and fixed L_g/d = 5 (gradients) cut-off frequency versus 1/L_{eff}. V_g corresponds to the peak g_m, and V_d = 2 V. 87

5.6 $InGaAs$ HEMTs with fixed d = 15 nm (triangles) and fixed L_g/d = 5 (gradients) cut-off frequency vs. L_g (solid line) and versus L_{eff} (dashed line). Vg corresponds to the peak gm, and V_d = 2 V. 87

Figure Page

5.7 *InGaAs* HEMT with d =15 *nm* (triangles), and *InGaAs* HEMT with a fixed aspect ratio of 5, and variable d, (gradients), average velocity in the channel region and L_{eff} for L_g = 30 *nm*. 88

5.8 *InGaAs* HEMT with d = 15 *nm* (triangles), and *GaN* HEMT with d = 15 *nm* (circles) average velocity in the channel region and L_{eff} for L_g = 75 *nm*. 89

5.9 LEFT: Electron velocity profiles for same gate length but reducing gate-to-channel distance (*i.e*. increasing L_g/d). RIGHT: f_T versus L_G^{-1}. . . 90

5.10 f_T versus L_{eff}^{-1} : this relation is linear regardless the adopted vertical geometry (*i.e*. value of d) and corresponds to the velocity characteristic of the material. 91

5.11 Separated scattering mechanism along the channel in linear (left) and logarithmic scale (right). 92

5.12 Transconductance (solid line) and f_T (dashed lines) for the *InGaAs* HEMT device L_{sg} scaling with L_g = 76 *nm*. 94

5.13 Average velocity in the channel region for the *InGaAs* HEMT device L_{sg} scaling with L_g = 76 *nm*. The origin of the axis corresponds to the source. 94

6.1 Generic schematic cross-section of the *GaN* HEMT simulation domain. . 99

6.2 Agreement between simulated [11] and experimental [12] DC (left) and RF (right) characterization. The experimental f_T is 153 *GHz*. 100

Figure Page

6.3 Comparison of the effective gate length extracted from the simulated electron velocity profile along the channel in the air (solid line) and *SiN* (dashed line) cases. 101

6.4 Velocity profile along the channel in the air (solid line) and *SiN* (dashed line) cases with L_{arm} = 500 *nm* and T-gate height of 80 and 130 *nm*. . . . 103

6.5 Electric field and velocity profile (dots) along the channel in the air (solid line) and *SiN* (dashed line) cases with L_{arm} = 500 *nm* and T-gate height of 80 *nm*.. 104

6.6 Electron energy profile along the channel in the air (solid line) and SiN (dashed line) cases with Larm = 500 nm and T-gate height of 80 nm. . . 106

6.7 Average carrier distribution density in the momentum space at a position along the channel corresponding to the electric field peak in a device with air. 107

6.8 Average carrier distribution density in the momentum space corresponding to different positions along the channel in a device with air.. 108

6.9 Electron scattering profile along the channel in the air (solid line) and SiN (dashed line) cases with L_{arm} = 500 nm and T-gate height of 80nm . 109

6.10 Electric field profile along the channel versus different T-gate arm lengths for the air (solid line) and SiN (dashed line) case with a T-gate height of 80 nm. 111

Figure Page

6.11 Electron scattering profile along the channel versus different T-gate arm lengths in the air (solid line) and SiN (dashed line) cases with a T-gate height of 80 nm. 112

6.12 Comparison between computational time required by the CMC (triangles) and EMC (circles) algorithm to simulate additional 6 ps of the bias point corresponding to the maximum f_{max}. 113

6.13 Air (left) and SiN (case) case EMC simulation plot of the different electron scatteringmechanism profiles along the channel with L_{arm} =500 nm and Tgate height of 80 nm.. 115

6.14 Electron density profile along the channel versus different T-gate arm lengths (L_{arm}) in the air (solid line) and SiN (dashed line) cases with a T-gate height of 80 nm. 117

6.15 Imaginary part of Y_{12} versus frequency for the air (solid line) and SiN (line with circles) case with L_{arm} = 500 nm and T-gate height of 80nm. . 118

6.16 Gate-to-drain feedback capacitance, extracted at 40 GHz through small-signal analysis from Im{Y_{12}}, in air (solid line) and SiN (dashed) cases with increasing Larm and T-gate height of 80 nm. 119

6.17 Impact of L_{arm} on the large-signal output impedance at 25 *GHz* in a de--vice with *SiN* dielectric constant and a T-gate height of 80 *nm*. 120

7.1 Small-signal equivalent circuit model of the intrinsic device and the extrinsic parasitic network model.. 125

7.2 Effect of the gate resistance on the Unilater Power Gain and Stability Factor. f_{max} versus R_G^{-1} is reported in the small inset in the left figure. . . 126

Figure Page

7.3 Effect of the source and drain inductance on the Short-circuit Current Gain (left) and the Stability Factor (right). 127

7.4 Example of the series impedance topology: FET with a series source resistance. 130

7.5 Example of the series impedance topology: resulting DC extrinsic drain current and transconductance versus gate voltage. 131

7.6 Example of the series impedance topology: resulting AC intrinsic source voltage. 132

7.7 Example of the series impedance topology: FET with a series drain resistor (left) and resulting drain current versus drain voltage (right). . . . 132

7.8 Example of the series impedance topology: FET with a series source inductor and resulting AC intrinsic source voltage versus time 133

7.9 Example of the bias tee topology: resistive load connected to the drain contact. 134

7.10 Example of the bias tee topology: output currents and voltages with an input gate voltage sine of 25 *GHz*. The device has width of 100 μ*m*. LEFT: $R = 5$ W, $L = 50$ *pH*. RIGHT: $R = 5$ W, $L = 500$ *pH*. 135

7.11 Moving average smoothing frequency response function between 0-100 *GHz* (left) and [0-600] *GHz* (right). 137

7.12 Down-sampling example. 139

7.13 Schematic example of the general HB algorithm: the source inductor is emulated by generating the voltage sinusoids at the different harmonics. 142

Figure Page

7.14 Simulation results of a source inductor emulated through HB by gene- -rating the voltage sinusoids for the first three harmonics. 143

7.15 Harmonic content of the analyzed source current, and of the voltage waveform generated by the Harmonic Balance algorithm. 144

7.16 The circuit including a diode excited by a large RF signal (a) can be divided into a pair of equivalent circuits, one describing the linear part (b), and another the non-linear part (c). 146

7.17 Series *RLC* resonant tank HB solver input mode. 156

7.18 Parallel *RLC* resonant tank HB solver input mode. 156

7.19 Manually-specified impedance HB solver input mode. 157

8.1 Example of a Class-B amplifier (*i.e.* conducting only for half of the input cycle) with high-Q matching network emulated by an active load-line technique. 161

8.2 Smith chart plot of the load tuning iterative procedure for a device biased for Class-A operations at 25 *GHz* with emulated load impedance seen by the device at its output contacts of 50 Ω. 162

8.3 Example of a Class-F amplifier (*i.e.* conducting only for half of the input cycle and harmonic loading) with high-Q matching network and different load impedance values at the different harmonics emulated by active load-lines [13, 14]. 164

8.4 EMC and CMC simulation time (1HB iteration) vs increasing field plate length 166

Figure Page

8.5 Example of a Class-B amplifier (*i.e.* conducting only for half of the input cycle) with input and output high-Q matching networks both of them emulated by an active load-line technique. 167

8.6 Example of a Class-B amplifier (*i.e.* conducting only for half of the input cycle) with high-Q matching network and parasitic gate resistan- -ce, R_G, both of them emulated by an active load-line technique. 169

8.7 Effects of the gate resistance, R_G, on the intrinsic gate voltage waveform of a FET used as Class-A RF power amplifier at 35 *GHz*. R_G was simulated by using both the FDTD/CMC and HB/CMC approach. . 170

8.8 Effect of the gate parasitic resistance, R_G, on the dynamic load-line at 100 *GHz*. 172

8.9 Effect of the source contact resistance R_{Sc} (left) and the drain contact resistance R_{Dc} (right) at 25 GHz. 173

8.10 Effect of source and drain contact resistance at 25 GHz with the device biased for Class-B operations. The overlaid DC I_D−V_D corresponds to the R_{Sc} = 0 Ω and R_{Dc} = 0 Ω case. 174

8.11 Effect of threading dislocation defects at 25 GHz with the device biased for Class-A operations. The overlaid DC I_D−V_D corresponds to a dislo- -cation density N_{dis} = 0 cm^{-3} case. 175

8.12 Effect of surface traps, located in both the source-gate and gate-drain regions (left) and only in the gate-drain region (right), at 25 GHz with the device biased for Class-A operations. The overlaid DC I_D−V_D corre- -sponds to a surface traps density N_{TR} = 0 cm^{-2} case. 176

Figure Page

8.13 Output current waveforms for increasing input power at 25 GHz in Class-B operations with high-Q output matching network and emulated load $Z_L = 50\ \Omega$. 177

8.14 Output current waveforms (left) and related dynamic load-lines (right) for increasing input power with the device biased for Class-A with high Q output matching network at 25 GHz. 178

8.15 Output power (P_{out}), Power Added Efficiency (PAE), and Gain (left) and drain current harmonic content (right) versus input power (P_{in}) at 25 GHz with the device biased for Class-A. 178

8.16 Impact of the field plate length (L_{ARM}) on the large-signal output impedance in a SiN passivated device at 25 GHz. 180

8.17 LEFT: Layout of the *InAlN/GaN* HEMT simulation domain and agreement between the DC simulated and experimental results [15]. RIGHT: AC small-signal Monte Carlo simulations. 182

8.18 Effects of the gate resistance, R_G, on the 35 *GHz* (left) and 70 *GHz* (right) Class-A RF power amplifier performance of the device in Figure 8.17. R_G was simulated by using the FDTD/CMC approach. 183

8.19 LEFT: Comparison between the FDTD and the HB circuit. RIGHT: RF power amplifier performance characterization at 35 *GHz* for the Class-A operations obtained by simulating R_{Sc} through FDTD and HB 184

8.20 Effects of the source contact series resistance, R_{Sc}, (left) and the drain contact series resistance, R_{Dc}, (right) on the 35 *GHz* Class-A RF power amplifier performance of the device in Figure 8.17. 185

Figure Page

8.21 Output power (P_{out}) measured at for different combination of the gate, R_G, and the source/drain, R_{Sc} and R_{Dc}, parasitic resistances. 186

8.22 Effects of threading dislocations (left) and surface traps (right) on the 35 *GHz* Class-A RF power amplifier performance of the device in Figu--re 8.17. 187

8.23 Output power (P_{out}) measured at for different combination of disloca--tion density (N_{DIS}) and surface trap density (N_T). 187

8.24 Comparison between the Class-A and Class-B RF power amplifier performance of the device in Figure 8.17. 188

Chapter 1

Introduction

High electron mobility transistors (HEMTs) represent a cutting-edge technology for high-power and high-frequency power amplifiers. Aerospace applications demand power amplifiers also capable of high-performance in terms of power efficiency, thermal dissipation, package dimensions, radiation hardness, reliability, and linearity. More specifically, space born communication systems commonly need to comply with stringent constraints of power, space aboard, and lifetime of the electronic components. Furthermore, high data-throughput multilevel modulation techniques that use non-constant envelope waveforms with a high Peak-to-Average Power Ratio (PAPR), result in stringent requirements in terms of the transistor linearity. On the other hand, such high-performance demand is leading to extreme large-signal operation in highly-scaled devices, resulting in highly non-linear, out-of-equilibrium transport of high-energy carriers inside the device [16, 17], which poses new challenges for the device engineering and modeling.

GaN HEMTs, based on wide band gap nitride materials, are very promising candidates for the next generation of high-power and high-frequency applications [2], such as radar and satellite communication front-ends with particular relevance for the design of state-of-the-art millimeter-wave power amplifiers in the Ka and X-band [18, 19], due to the combination of high electron velocity and break down voltage as reported in Figure 1.1. As epitaxial growth techniques have progressed, *GaN*-based devices have been widely investigated, and a much larger output power when compared with *GaAs*-based devices has been demonstrated [20] in the late 1990s. Such impressive performance is due to the high transconductance, the good thermal management, the high breakdown voltage, the excellent transport characteristics at high fields, high 2D electron gas (2DEG) density in the channel, high carrier confinement, and the high electron mobility. State-of-the-art devices have been reported with a small

	Si	GaAs	4H-SiC	GaN	Diamond
E_g (eV)	1.1	1.42	3.26	3.39	5.45
n_i (cm^{-3})	1.5×10^{10}	1.5×10^{6}	8.2×10^{-9}	1.9×10^{-10}	1.6×10^{-27}
ε_r	11.8	13.1	10	9.0	5.5
μ_n (cm^2/Vs)	1350	8500	700	1200(Bulk) 2000(2DEG)	1900
v_{sat} (10^7cm/s)	1.0	1.0	2.0	2.5	2.7
E_{br} (MV/cm)	0.3	0.4	3.0	3.3	5.6
Θ (W/cm K)	1.5	0.43	3.3-4.5	1.3	20
$JM = \frac{E_{br} v_{sat}}{2\pi}$	1	2.7	20	27.5	50

Figure 1.1: Material properties important for the device high-frequency, high-power performance [2]. The parameter JM is the Johnson's figure of merit [3].

signal current-gain cut-off frequency, f_T, of 245 GHz [15], power-gain cut-off frequency, f_{max}, of 300 GHz [21], and record output power of 40 W/mm [22]. This performance is constantly improving through the use of emerging device structures such as N-face [23, 24, 25, 4], $InGaN$ back-barrier [12, 26, 11, 27], and $AlInN$ lattice-matched barrier [15, 16] GaN HEMTs. Specifically, the wide band gap makes the nitride materials particularly suited for operations at elevated temperatures because they become intrinsic at much higher temperature than the narrow band gap materials. Moreover, a very large energy is required to initiate avalanche breakdown (*i.e.* promoting an electron from the valance band to the conduction band due to impact ionization), resulting in high breakdown fields and allowing devices based on these materials to sustain high currents and voltage levels. Furthermore, $AlGaN/GaN$ heterostructures show interesting properties due to large spontaneous and piezoelectric polarization effects [9]. Free carriers are induced by a net polarization charge at the heterostructure interface, creating a spontaneous electron channel without requiring any modulation doping unlike devices based on $AlGaAs/GaAs$ heterostructures.

On the other hand, for low-power and very high-frequency applications (such as next-generation digital circuits, sub-millimeter-wave communications, and terahertz electronics) HEMTs based on narrow band gap $InGaAs$ alloys are the best candidates [28, 29, 25, 30, 31]. The $InGaAs$ small effective mass and very high electron

mobility at room temperature (exceeding 10000 cm^2/Vs) make possible devices with high electron velocity in the channel. This, combined with ultra-short gate length, allows very fast-switching devices due to a small transit time in the gate region. f_t of 644 GHz and f_{max} of 681 THz simultaneously [32], and absolute f_{max} of 1.2 THz [33] were reported for state-of-the-art ultra-short *InGaAs* devices.

The need for demanding circuit performance required by present day applications is also leading to an enormous push forward of the transistor operating frequencies and power levels. In particular, improved frequency performance is achieved through aggressive scaling, resulting in nano-scale devices [25, 34, 31]. Moreover, large output power can be achieved with large current and voltage swing, which requires high-voltage biasing (exceeding 100 V for *GaN* RF-devices [2]), that drives the devices to a highly non-linear regime and also results in high electric fields close to the breakdown value. Under such conditions, highly non-linear out-of-equilibrium carrier transport occurs, and a more detailed understanding of how the nanoscale dynamics of high energy carriers affects the overall device performance is required to further improve the state-of-the-art of device engineering. In this context, the role of simulation tools is to provide this understanding, and to assist device designers with indications regarding the optimal device geometry and material choices instead of a costly trial-and-error prototyping approach. In particular, the Monte Carlo technique [35] allows an accurate description of the complex nanoscale carrier dynamics, and it is of great importance in providing an accurate picture of the inner physics of the device as it is described in the next paragraph.

1.1 The Cellular Monte Carlo Technique

The properties of the charge distribution in semiconductors are described, within the semi-classical approach, by the single particle *distribution function*, $f(\vec{r},\vec{k},t)$. In particular, $f(\vec{r},\vec{k},t)$ gives the classical probability of finding a carrier at location $\vec{r}$, with crystal momentum $\vec{k}$, at time t. Thus, $f(\vec{r},\vec{k},t)$ provides a description of the distribu-

tion of carriers both in position and momentum, and it can be used to obtain various quantities of interest such as the mean energy of the carriers or their drift velocity, as a function of the electric field, lattice temperature, and/or carrier concentration gradient. The time evolution of $f(\vec{r},\vec{k},t)$ is described by the Boltzmann Transport Equation (BTE) [36]:

$$\frac{\partial f}{\partial t} = -\vec{v}\cdot\nabla_{\vec{r}} f - \dot{\vec{k}}\cdot\nabla_{\vec{k}} f + \left(\frac{\partial f}{\partial t}\right)_{coll}, \tag{1.1}$$

where $\vec{v}$ is the carrier velocity and the last term on the right hand is the rate of change of f due to collisions.

The distribution function can be obtained by solving the BTE but, however the BTE is a multi-dimensional non-linear differential equation, solving this equation directly is still an open issue. Approximations of the BTE can be done in order to simplify the equation to a directly resolvable form. In particular, approximations based on the first two moments of the BTE yield, respectively, the drift-diffusion and hydrodynamic models [37]. However, these simplified models require reliable transport parameters, such as energy relaxation times or mobility, and their validity domain is limited.

Alternatively, a space-time statistical solution of the semi-classical BTE can be obtained through the particle-based Ensemble Monte Carlo (EMC) method [35]. Within this EMC framework, the distribution function is not obtained solving analytically the BTE based on some sort of approximation, but it is rather obtained actually simulating the dynamics of an ensemble of carriers, tracked both in the real and the momentum space, under the effect of electric fields and scattering processes. Then, quantities such as the average energy, drift velocity, or the mobility can be obtained by averaging over the ensemble of carriers.

The EMC method has been proved successful and more accurate than traditional drift-diffusion methods [38] without the need of relying on the accuracy of transport parameters, which are often an output of the EMC simulation itself. Moreover, it

allows a better modeling of high-energy non-stationary transport phenomena occurring in high-field regimes, where some of the drift-diffusion approximations may lose their validity.

Scattering rates are pre-calculated and tabulated combining material characteristics such as the electronic band structure, phonon spectra, doping density, dislocation density, and piezoelectric coefficients. The initial carrier distribution inside the computational domain representing the semiconductor device is computed by imposing the neutrality condition on mobile carriers, fixed ions, and other fixed charges. Then, the real-time carrier distribution is used to solve the Poisson's equation and extract the electric field distribution. The field is kept constant for a short ballistic free-flight time step and is used to update the carrier positions in the momentum space. The duration of the time-step in which the field is considered constant is crucial in order to prevent numerical artifacts and plasma oscillations [38]. At the end of each free flight, a stochastic Monte Carlo procedure selects whether a scattering event occurs and, if this is the case, the exact scattering mechanisms based on the probability tabulated during the initialization setup. The carrier position in the momentum space is then recomputed after each scattering. Then, the same sequence is iterated, with typically 4-5 free-flight steps for each field update step, until the simulation achieves the desired steady-state.

The Cellular Monte Carlo (CMC) algorithm [39] was developed to reduce the high computational demand of the traditional EMC approach, due to the recalculation of the positions of the carriers in the momentum space required when a scattering event occurs at the end of a free flight. The CMC approach differs from the conventional EMC method in terms of selection of the final state. In the traditional EMC approach, a search of the whole discretized BZ has to be performed after each scattering event in order to invert the energy/momentum relation, for finding all the energy conserving cells in the momentum space and then the final state. On the other hand, the CMC method involves the use of a pre-computated transition look-up table for the total probability

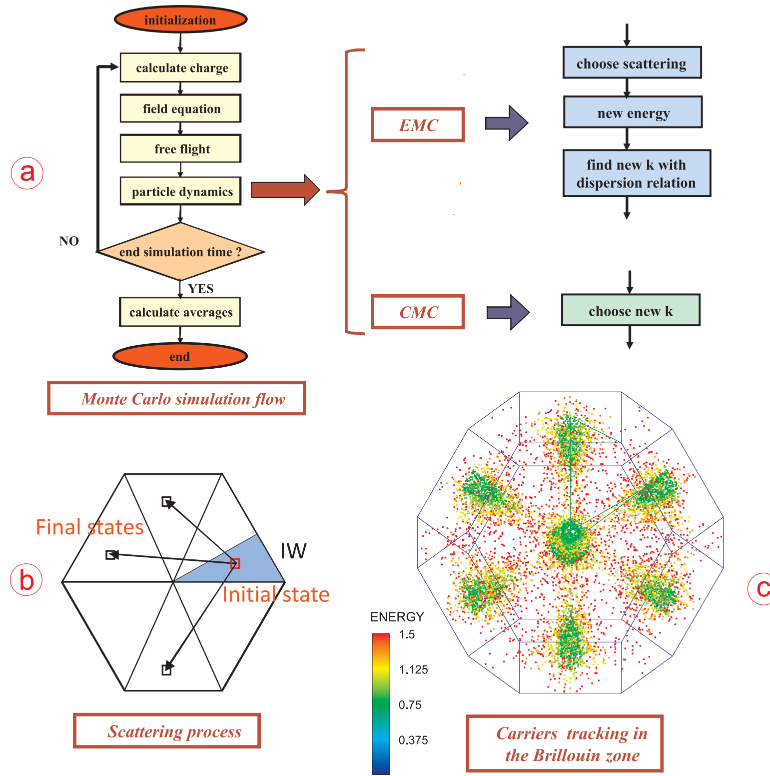

Figure 1.2: Monte Carlo simulation: a) Illustration of the simulation flow and difference between EMC/CMC algorithms. b) Example of carrier initial state in the momentum space, and possible final states to be chosen during the scattering process. c) Particle tracking in the momentum space within the first Brillouin zone.

of scattering, due to all mechanisms, from every initial state to every final state. In this way, the selection of the final state is reduced to the generation of a single random number, which results in a very efficient simulation of a large number of scattering events. This allows to efficiently simulate the high-energy carrier transport due to the large electric field occurring in state-of-the-art devices operating as high-power amplifiers. On the other hand, the price for this reduction of the scattering process calculation is the size of the look-up table, which often requires more than three gigabytes of RAM. Moreover, by storing only the total rate, information is lost about the exact type of scattering process involved in the transition (e.g., the nature of the phonon involved, and its energy). Thus, the much less performing EMC (up to 25 times slower [39]) may be still useful when obtaining a plot of the separate scattering mechanisms is of interest. In

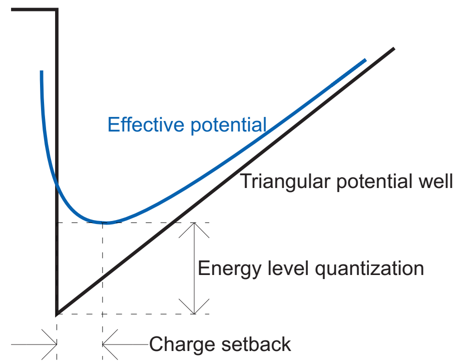

Figure 1.3: Quantum mechanical confinement effects (charge setback and energy level quantization) in a triangular potential well obtained through effective potential approach.

any case, the discretization of the first Brillouin zone (BZ1) of the crystal [40] into an inhomogeneous tensor-product grid, and the exploitation of the symmetry of the crystal such that the table is computed only for initial states in the irreducible wedge [41] of the BZ1, allows an efficient use of the memory and consequent minimization of the transition table memory requirement.

Our Cellular Monte Carlo code includes scattering processes due to polar optical phonons, acoustic and optical mode deformation potential scattering, piezoelectric scattering, ionized impurity, and dislocation scattering within a fullband framework. The fullband electronic structure across the entire BZ is obtained through nonlocal empirical pseudopotential methods (EPM) [42, 43]. Details regarding the adopted fullband phonon spectra and band structure for wurtzite *GaN* used in the present work can be found in [44, 45]. A particle-based dynamics kernel self-consistently coupled with a multi-grid Poisson solver [46] is used to calculate charge distribution and track particles in the real and momentum space.

Quantum mechanical effects in the channel region, due to confinement of the 2DEG at the heterojunction interface, are accounted by using an effective potential approach [47]. In particular, this method allows to reproduce quantum effects such as charge setback and energy level quantization [48] as shown in Figure 1.3.

1.2 Purpose of the Work

The purpose of the present work is to investigate and provide a meaningful insight about nanoscale physics of state-of-the-art HEMTs through the Cellular Monte Carlo technique. In particular, emphasis is given to the connection between nanoscale carrier dynamics and device performance, with particular focus on the modeling of high-power devices for millimeter-wave power amplifier applications, sub-millimeter band devices, and also Terahertz electronics. With this scope, HEMTs based on novel materials and structures are simulated, and their DC/frequency performance is reported and explained in order to provide useful design guidelines for device engineers. In particular, a specific effort is made to link device characterization to the parameters relevant for transistor circuit models as well, allowing circuit designers to properly choose operational conditions such as frequency, biasing, and power handling. Furthermore, efficient simulation techniques that allow the large-signal characterization of HEMTs used as power amplifiers in the mm and sub-mm band are also introduced, and simulations of the non-linear effects arising in FETs under large-signal operations are carried out.

We first discuss the emerging *N*-face technology in Chapter 3 by introducing the theory of polarization in nitride materials, and by analyzing the effects of device scaling on the DC and RF performance of *N*-face *GaN* HEMTs. In Chapter 4, a direct comparison between devices based on this new technology and conventional *Ga*-face HEMTs is done by obtaining the analogous *Ga*-face device starting from a *N*-face device layout. In particular, a modeling approach is devised to allow a fair comparison between the two technologies. Next, in Chapter 5, the short-channels effects and the vertical and lateral scaling are investigated in *GaN* and *InGaAs* HEMTs, with particular focus on the differences between devices based on these two material families. The role of passivation dielectric and gate geometry on the RF performance of millimeter-wave

power *GaN* HEMTs is discussed in Chapter 6, where small-signal equivalent circuit parameters are extracted and discussed as well. In Chapter 7, different techniques are introduced for performing circuit-device simulations capable of accounting for the interaction between the device and the external circuitry (*i.e.* matching networks and parasitic elements) occurring during small-signal and large-signal operations. More specifically, an efficient large-signal circuit-device simulation approach, consisting of a Harmonic Balance frequency-domain circuit solver self-consistently coupled with our Monte Carlo particle-based device simulator, is discussed in depth. Such techniques are then applied in Chapter 8, where the large-signal characterization of state-of-the-art *GaN* HEMT power amplifiers, including the highly frequency selective matching network, is performed for the first time with a Monte Carlo code.

Chapter 2

High-frequency Characterization of Millimeter and Submillimeter-band FET Amplifiers

2.1 Introduction

Several figures of merit are often used to assess the performance of microwave devices. The most important metrics related to the high-frequency performance of microwave transistors are the short-circuit current gain cut-off frequency, f_T, and the maximum oscillation frequency, f_{max}. The short-circuit current gain, h_{21}, is defined as the ratio of the small-signal output current to the small-signal input current of the device when the output terminals are shorted, and f_t is the frequency where this gain drops to the unity. Similarly, f_{max} corresponds to the unity power gain frequency.

2.2 FET Two-port Network Model

The most common way to characterize the small-signal behavior of Field Effect Transistors (FETs) is to model the device as a two-port network as shown in Figure 2.1. This model is fully described by the frequency-dependent Y-parameters [49], which define the relations between currents and voltages as follows:

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}. \tag{2.1}$$

In a common source configuration, the subscript 1 refers to the port between the gate and the source, while the subscript 2 refers to the port between the drain and the source. Following this notation, the Y-parameters are defined as follows:

$$y_{11} \equiv \left. \frac{i_1}{v_1} \right|_{v_2=0}, \tag{2.2}$$

$$y_{12} \equiv \left. \frac{i_1}{v_2} \right|_{v_1=0}, \tag{2.3}$$

$$y_{21} \equiv \left. \frac{i_2}{v_1} \right|_{v_2=0}, \tag{2.4}$$

$$y_{22} \equiv \left. \frac{i_2}{v_2} \right|_{v_1=0}. \tag{2.5}$$

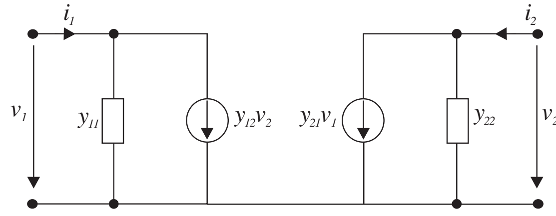

Figure 2.1: Equivalent two-port network of the FET based on the Y-parameters.

Using the Fourier Series expansion, the input and output signals can be expressed as a sum of sinusoids of different amplitudes and frequencies. Therefore, one can introduce the frequency-dependent Y-parameters as follows:

$$\begin{bmatrix} \tilde{i_1}(\omega) \\ \tilde{i_2}(\omega) \end{bmatrix} = \begin{bmatrix} Y_{11}(\omega) & Y_{12}(\omega) \\ Y_{21}(\omega) & Y_{22}(\omega) \end{bmatrix} \begin{bmatrix} \tilde{v}_1(\omega) \\ \tilde{v}_2(\omega) \end{bmatrix}, \tag{2.6}$$

where the $(\tilde{x})$ symbol denotes a small variation about the steady-state (e.g. $\tilde{v}_1(\omega)$ is the amplitude of the small voltage variations at frequency ω over a DC bias).

The Y parameters are then defined in the frequency domain as follows:

$$Y_{11}(\omega) = \frac{\tilde{i_g}(\omega)}{\tilde{v}_{gs}(\omega)}\Big|_{\tilde{v}_{ds}=0}, \tag{2.7}$$

$$Y_{12}(\omega) = \frac{\tilde{i_g}(\omega)}{\tilde{v}_{ds}(\omega)}\Big|_{\tilde{v}_{gs}=0}, \tag{2.8}$$

$$Y_{21}(\omega) = \frac{\tilde{i_d}(\omega)}{\tilde{v}_{gs}(\omega)}\Big|_{\tilde{v}_{ds}=0}, \tag{2.9}$$

$$Y_{22}(\omega) = \frac{\tilde{i_d}(\omega)}{\tilde{v}_{ds}(\omega)}\Big|_{\tilde{v}_{gs}=0}, \tag{2.10}$$

where the subscripts s, g, and d refer to the source, gate, and drain terminal, respectively.

The Y-parameters can be obtained through the Fourier decomposition (FD) method [50, 51, 52, 53], a technique where a small step voltage perturbation is applied to one electrode of the device in steady state at the chosen bias point. The resulting output response is Fourier analyzed to obtain the full frequency response of the device modeled as a two-port network. With such a technique, small-signal parameters can be

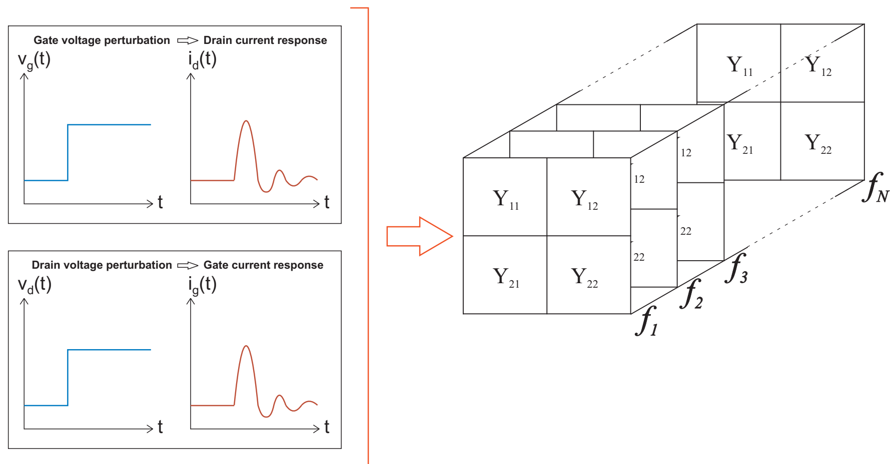

Figure 2.2: Full-frequency Y-parameter extraction through step perturbations applied through two separate simulations.

extracted over a wide frequency range without any need for sweeping each frequency singularly. The complete set of Y-parameters is obtained by applying a step perturbation, with two separate simulations, to the gate and to the drain, as shown in Figure 2.2. In particular, applying a voltage step perturbation at one port while keeping the voltage fixed at the other port, where the output current response is measured, is equivalent to an AC short-circuit at this output port. This is because a situation where a non-zero output AC current is associated to a zero output AC voltage corresponds to an output port load-line with impedance equal to zero. Thus, the conditions required by Eq. 2.2- 2.5 are simply satisfied by applying a voltage perturbation at the input port and measuring the current at the output port where a fixed DC voltage is applied. The Y-parameter for a specific frequency value can be also obtained by simply applying a sinusoidal perturbation instead the step one. The applied perturbations must be small enough to preserve the device linear response, avoiding in such way the generation of harmonics and intermodulation products typical of non-linear responses (see Section 2.5), but large enough to allow a good signal-to-noise ratio of the extraction process.

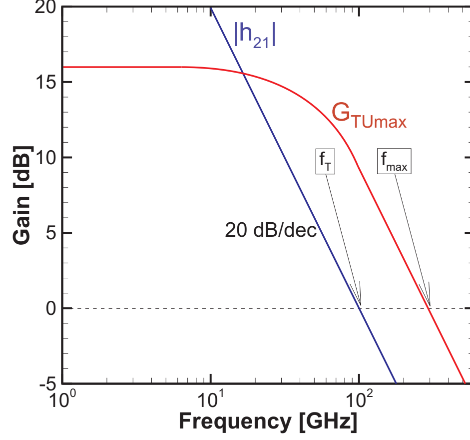

Figure 2.3: Estimation of f_T and f_{max} from, respectively, the 0 dB intercept point of the short-circuit current gain $h_{21}(f)$ and the maximum unilateral transducer gain, $G_{TUmax}(f)$.

Once the Y-parameters are known over a wide frequency range, we can extract f_T from the 0 dB intercept point of the the short-circuit current gain $h_{21}(f)$:

$$h_{21} = \frac{Y_{21}}{Y_{11}} = \frac{i_d}{i_g}. \tag{2.11}$$

It should be noted, by observing the definition of h_{21}, that the current gain does not necessary drop with increasing frequency due to a decrease of the output current i_d. The increase of i_g with increasing frequency, flowing through the device input capacitance, can usually play a major role. As general trend, h_{21} is expected to have a 20 dB/dec slope and to have a trend toward infinite at DC (*i.e.* 0 frequency) in FETs. Similarly, f_{max} can be extracted from the 0 dB intercept point of the maximum unilateral transducer gain, $G_{TU,max}(f)$, which corresponds to the gain that would be achieved if the transistor were unilateralized (*i.e.* internal feedback neutralized through an external lossless network) and conjugately matched with the source and load impedance:

$$G_{TUmax} = \frac{|Y_{21} - Y_{12}|^2}{4[Re(Y_{11})Re(Y_{22}) - Re(Y_{12})Re(Y_{21})]}. \tag{2.12}$$

Furthermore, the Z-parameters and the S-parameters [54] can be easily obtained through matrix operations:

$$\mathbf{Z} = \mathbf{Y}^{-1}, \tag{2.13}$$

and

$$\begin{aligned} \mathbf{S} &= -(\mathbf{Y}+\mathbf{Y}_0)^{-1}(\mathbf{Y}-\mathbf{Y}_0) \\ &= -(\mathbf{Y}-\mathbf{Y}_0)(\mathbf{Y}+\mathbf{Y}_0)^{-1}, \end{aligned} \tag{2.14}$$

where Y_0 is a diagonal matrix with the characteristic admittance on its diagonal.

Alternatively, the S-parameters can be obtained directly from the Z-parameters:

$$\begin{aligned} \mathbf{S} &= (\mathbf{Z}+\mathbf{Z}_0)^{-1}(\mathbf{Z}-\mathbf{Z}_0) \\ &= (\mathbf{Z}-\mathbf{Z}_0)(\mathbf{Z}+\mathbf{Z}_0)^{-1}, \end{aligned} \tag{2.15}$$

where Z_0 is a diagonal matrix with the characteristic impedance on its diagonal.

Physical Meaning of f_T

An alternative way to extract f_T can be obtained directly through its physical meaning in terms of carrier dynamics. The short-circuit current-gain cut-off frequency can be related to the carrier transit time under the region controlled by the gate, and expressed in terms of the average velocity and effective gate length [55]:

$$\tau_T = \frac{1}{2\pi f_T} = \int_{L_{eff}} \frac{1}{v_{ave}(x)} dx, \tag{2.16}$$

where L_{eff} is the effective gate length, and $v_{ave}(x)$ is the velocity calculated along the channel (x-direction) and is obtained as a weighted average along the y-direction with respect to the carrier density.

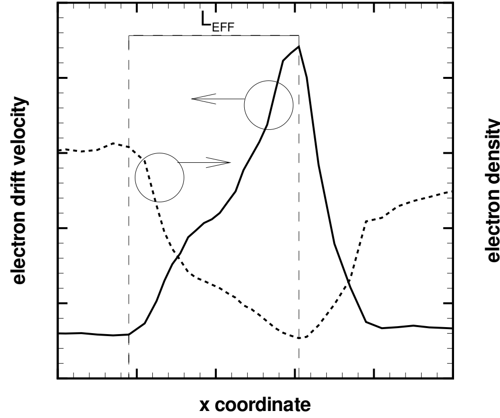

Figure 2.4: Schematic example of effective gate length measurement based on carrier velocity (solid line) and density (dashed line) profiles along the channel.

This is just a more sophisticated version of the common relation between f_T and carrier velocity usually found in the textbooks [56]:

$$f_T = \frac{v}{2\pi L} = \frac{1}{2\pi\tau_T}, \tag{2.17}$$

where v is the carrier velocity and L is the gate length.

The gate extends its control (*i.e.* the ability of modulating the carrier density) over a region longer than the gate metallurgical extension (L_g) due to fringing capacitances, the length of such region is called the effective gate length. Thus, the transit time is critical in term of maximum frequency of an oscillating signal applied to the gate that modulates the electron density and should allow modulated electrons, located at the source-end of the gate, to fully transit along the gate control region (*i.e.* L_{eff}). In the negative case, a large displacement gate current will be the result. In our work, we followed the L_{eff} definition given in [55], where L_{eff} is measured from the electron acceleration starting point to the velocity peak. Moreover, the beginning of the depletion region under the influence of gate corresponds to the electron acceleration starting point as well. An example of the L_{eff} measurement according to this definition, in relation to the electron velocity and density profile along the channel, can be seen in

Figure 2.4. Once the velocity profile is obtained from DC-analysis and the extension of L_{eff} is established, we can obtain a direct estimation of f_T by using eq.(2.16).

Thus, we can validate our estimation of f_t through an independent cross-comparison by adopting this method and the one described in the previous section. In particular, deciding in an exact way the extension of L_{eff} is not always a trivial task, and it is the critical factor for a reliable calculation of the transit time. On the other hand, the critical task for the method based on small-signal analysis and Fourier decomposition is the determination of an optimal Fourier transform window length during the post-processing extraction. The comparison of the values obtained with both methods reduces the indetermination of the f_T quantitative estimates.

2.3 Small-signal FET Equivalent AC Circuit

The extraction of small-signal equivalent circuit parameters is of great importance for accurate transistor compact models required by microwave circuit designers. An example of a typical FET small-signal equivalent circuit based on the models described in [53, 57] can be seen in Figure 2.5. In our case, the ”intrinsic device” modeled by the Monte Carlo simulator corresponds to the intrinsic gate area plus the source-to-gate and drain-to-gate access regions. This intrinsic device is simulated by the CMC code with minimal assumptions on the device behavior by representing the nanoscale physics and the carrier dynamics. Thus, the device can be seen as a ”black box” with intrinsic voltages as input and intrinsic currents as output. In the framework of small-signal analysis, the intrinsic parameters in the model can be associated to the device internal nanoscale description (*i.e.* the evolution of the carrier distribution in space and time). Therefore, the intrinsic parameters are closely related to the nanoscale physics of the device, and can be accurately estimated with our fullband Monte Carlo simulator output. However, these intrinsic parameters are bias-dependent because the device internal description depends on the applied DC bias.

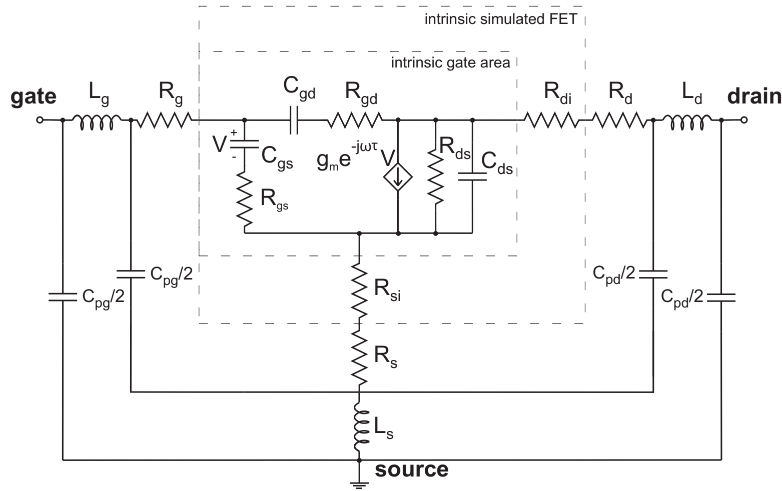

Figure 2.5: FET small-signal equivalent circuit including extrinsic parasitic components.

On the other hand, the extrinsic parasitic parameters are related to external passive structures (*i.e.* interconnections capacitive/inductive couplings and resistance), and they are not in general bias-dependent being passive and not related to the device. In particular, the extrinsic inductive elements take into account inductive coupling due to wire bonding, while capacitors are related to pad coupling, and resistive elements to conductor Ohmic resistances. The knowledge of these external parameters is usually experimentally available through cold-extraction techniques [58, 59]. In this approach, measurements are performed on the FET with a specific bias (*i.e.* zero bias at each contact) that allows simplifying the equivalent circuit. In such way, bias-independent extrinsic parameters can be isolated and then directly measured. The intrinsic parameters can be also experimentally determined by de-embedding the extrinsic parameters from the small-signal characterization of the device [60].

The small-signal equivalent circuit intrinsic parameters can be easily obtained from the two-port parameters of the simulated intrinsic device once an equivalent circuit is chosen. For instance, this can be done by discriminating which equivalent circuit parameters contribute to the input admittance with output short-circuit (*i.e.* Y_{11}), output admittance with input short-circuit (*i.e.* Y_{22}), forward transconductance with output

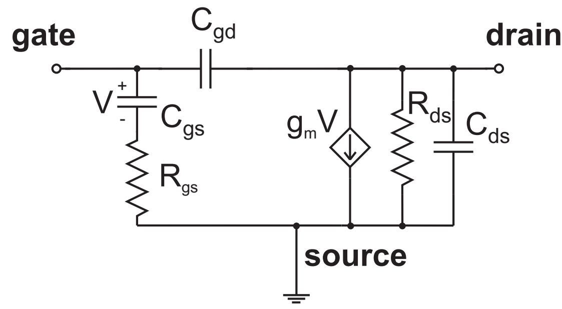

Figure 2.6: Simplified intrinsic FET small-signal equivalent circuit.

short-circuit (*i.e.* Y_{21}), and reverse transconductance with input short-circuit (*i.e.* Y_{12}). In particular, the accuracy of the extraction of these parameters depends on the complexity of the adopted models. However, a more complex model with several equivalent circuit elements does not always result in an improved accuracy if the elements to be included are not properly chosen. For example, including an element that does not represent any physical mechanism in the device will negatively affect the extraction of the other parameters. Moreover, some reactive components may or may not be relevant depending on the frequency of interest. Thus, a simple model can still be reliable and, within a certain frequency range, sufficiently accurate. As an example, the relation between the two-port parameters and the circuit parameters related to the simplified intrinsic FET equivalent circuit shown in Figure 2.6 can be best established through the Y-parameters. Starting from the definition of the Y-parameters:

$$Y_{11} = \frac{j\omega C_{gs}}{1 + j\omega R_{gs} C_{gs}} + j\omega C_{gd}, \tag{2.18}$$

$$Y_{21} = \frac{g_m}{1 + j\omega R_{gs} C_{gs}} - j\omega C_{gd}, \tag{2.19}$$

$$Y_{12} = -j\omega C_{gd}, \tag{2.20}$$

$$Y_{22} = \frac{1}{R_{ds}} + j\omega (C_{ds} + C_{gd}). \tag{2.21}$$

An example of the estimation of C_{gd} based on the imaginary part of Y_{12}, extracted from a step perturbation post-processed with several different Fourier transform windows, can be seen in Figure 2.7. Moreover, the small-signal equivalent circuit parameters

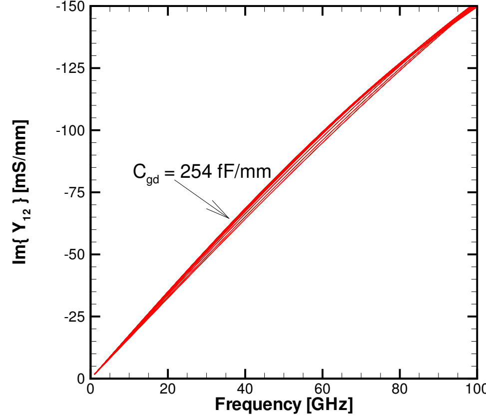

Figure 2.7: Estimation of C_{gd} from the the imaginary part of Y_{12} obtained by step perturbation. The spreading of the different curves is due to the different windows used for the Fourier transform.

can be also related to the device DC/AC performance and extracted from the device characterization [56]:

$$f_T \approx \frac{g_m}{2\pi C_g} = \frac{g_m}{2\pi (C_{gs} + C_{gd})}, \tag{2.22}$$

$$f_{max} \approx \sqrt{\frac{f_T}{8\pi R_g C_{gd}}}. \tag{2.23}$$

With a similar approach, experimentally measured extrinsic parasitic parameters can be analytically embedded in the small-signal characterization of our simulated intrinsic device through a post-processing operation [53]. This is accomplish by simply deriving the two-port model of the parasitic element of interest in the form of Y or Z-matrix , then this extrinsic parameter matrix is added to the intrinsic device matrix [54] as described in Chapter 7. The Y or Z-matrix form to use depends on whether the parasitic can be considered a shunt or series network with respect to the device.

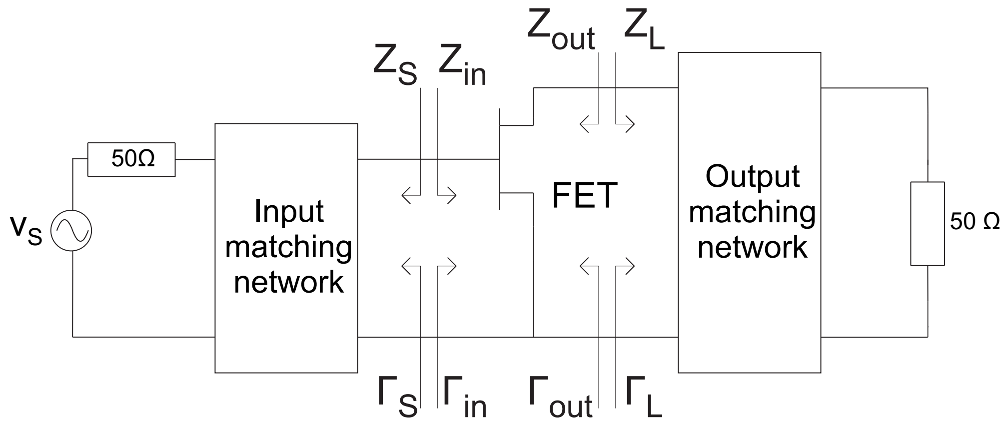

Figure 2.8: Microwave amplifier block diagram.

2.4 FET RF Small-signal Amplifiers

Once the complete two-port network parametrization described is Section 2.2 is known, the device small-signal amplifier performance can be completely assessed in an analytical way. The block diagram of a typical RF amplifier that exploits the analytical formalism described in this section is shown in Figure 2.8. A more rigorous description of the formalism described in this section can be found in the Gonzalez book [54], and the Pozar book [49].

The transistor input and output impedance, as well as the source generator and load impedance, are usually expressed in terms of reflection coefficients with respect to a reference impedance, Z_0, which usually corresponds to the transmission line characteristic impedance (even though distributed effects are not directly included within this formalism):

$$\Gamma_{in} = \frac{Z_{in} - Z_0}{Z_{in} + Z_0}, \tag{2.24}$$

$$\Gamma_{out} = \frac{Z_{out} - Z_0}{Z_{out} + Z_0}, \tag{2.25}$$

$$\Gamma_S = \frac{Z_S - Z_0}{Z_S + Z_0}, \tag{2.26}$$

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0}, \tag{2.27}$$

where the input and output reflection coefficients can be directly obtained from the two-port network parametrization and the knowledge of the chosen Z_S and Z_L:

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}, \tag{2.28}$$

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S}. \tag{2.29}$$

As we can see, the input reflection coefficient (*i.e.* the transistor input impedance) is in general also dependent on what happens at the device output (through the device reverse transmission coefficient S_{12}). Conversely, the output reflection coefficient (*i.e.* the transistor output impedance) is in general also dependent on what happens at the device input (through the forward transmission coefficient S_{21}).

The lossless input and output matching networks transform the conventional 50 Ω impedance, associated to the source generator impedance and the load impedance, to the desired source and load impedance, Z_S (*i.e.* Γ_S) and Z_L (*i.e.* Γ_L), required to achieve the performance required by the design specifications. The impedance matching is achieved by changing the load current and voltage relative amplitude and phase relation seen by the device at its contacts.

The transducer gain can be expressed as:

$$G_T = \frac{1 - |\Gamma_S|^2}{|1 - \Gamma_{in}\Gamma_S|^2} \cdot |S_{21}|^2 \cdot \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2}, \tag{2.30}$$

or equivalently, as:

$$G_T = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} \cdot |S_{21}|^2 \cdot \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_{out}\Gamma_L|^2}. \tag{2.31}$$

Observing the two G_T expressions, we can see that in both cases the left most product term (which is related to the input side of the amplifier) and the right most term (which is related to the output side of the amplifier), are interdependent through the definitions of Γ_{in} and Γ_{out} (*i.e.* Γ_{in} depends on Γ_L, while Γ_{out} depends on Γ_S). Moreover, the optimum source and load impedance that maximize the power are given by the maximum

power transfer theorem stating that:

$$\begin{cases} Z_S = Z_{in}^* \\ Z_L = Z_{out}^* \end{cases} \tag{2.32}$$

which corresponds to:

$$\begin{cases} \Gamma_S & = \Gamma_{in}^* \\ \Gamma_L & = \Gamma_{out}^* \end{cases} \tag{2.33}$$

However, this condition also results in an interdependence between Γ_S and Γ_L for the reasons discussed before. In such a situation, for a given set of S-parameters, the values of Γ_S and Γ_L that simultaneously maximize the gain through a simultaneous conjugate match can be found through specific equations that take into account all these constraints [54].

In most cases, the transistor exhibits a weak reverse transmission coefficient ($i.e.$ $S_{12} << 1$) due to a small intrinsic feedback, or, alternatively, a large intrinsic feedback can be neutralized by an external lossless compensation network ($i.e.$ a feedback external inductor to cancel out the intrinsic drain-to-gate feedback capacitance in a FET). Under such conditions, the amplifier can be approximated as *unilateral* because the interdependence between input and output is eliminated.

The error of this approximation can be evaluated as:

$$\frac{1}{(1+U)^2} < \frac{G_T}{G_{TU}} < \frac{1}{(1-U)^2}, \tag{2.34}$$

where

$$U = \frac{|S_{12}||S_{21}||S_{11}||S_{22}|}{(1-|S_{11}|^2)(1-|S_{22}|^2)}, \tag{2.35}$$

is known as the *Unilateral Figure of Merit*.

Within the *unilateral approximation*:

$$\Gamma_{in}|_{S_{12}=0} = S_{11}, \quad (2.36)$$

$$\Gamma_{out}|_{S_{12}=0} = S_{22}, \quad (2.37)$$

resulting in the *Unilateral Transducer Gain* expression:

$$G_{TU} = \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2} \cdot |S_{21}|^2 \cdot \frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2}, \quad (2.38)$$

where the most left and the most right product terms are now independent, allowing to optimize G_{TU} by independently adjusting the source and load impedance. The gain can be maximized by applying the maximum power transfer theorem within the unilateral approximation:

$$\begin{cases} \Gamma_S & = \Gamma_{in}^* = S_{11}^* \\ \Gamma_L & = \Gamma_{out}^* = S_{22}^* \end{cases} \quad (2.39)$$

resulting in:

$$G_{TUmax} = \frac{1}{|1-|S_{11}|^2} \cdot |S_{21}|^2 \cdot \frac{1}{1-|S_{22}|^2}. \quad (2.40)$$

Note that, in general, the conjugate matching for maximum power transfer does not correspond to the transmission line impedance matching for no reflections.

It can been shown [54] that, while there is an unique impedance value (different for source and load impedance) that maximizes the gain, the impedance value associated with a sub-maximum gain is not unique. In particular, the locus of the Γ_L values, or conversely the Γ_S values, associated to a certain constant G_{TU} lies on a circle when plot in the Smith chart. In such way, it is possible to obtain the so called *Constant Gain Contours*, shown in Figure 2.9, where each circle corresponds to the locus related to a certain gain value.

The most inner point corresponds to the impedance that maximizes the gain (unique value), while the center of the other gain contour circles are aligned along

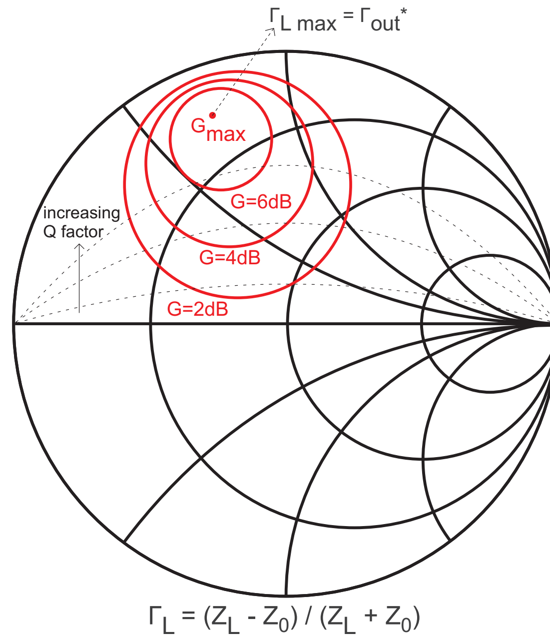

Figure 2.9: Typical small-signal amplifier constant gain, G, contour plots for different load reflection coefficients (Γ_L) in the Smith chart. The curves associated to a constant Q factor, associated to a certain lossless matching network required to convert a 50 Ω impedance (if $Z_0 = 50\ \Omega$) to the chosen load impedance Z_L, are also shown. The reference impedance is Z_0.

the line connecting the optimum impedance to the center of the Smith chart. These constant gain circles can be used to choose a sub-optimal source and load impedance in combination with the contour circles of other relevant figures of merit, derived in a similar fashion, when a trade-off is needed between gain and other figures of merit such as noise figure (NF), Voltage Standing Wave Ratio ($VSWR$), matching network Q-factors (*i.e.* narrow-band or broad-band matching), and input/output stability regions. Then, the chosen load or source impedance value can be obtained as:

$$Z_{L,S} = \left(\frac{1 + \Gamma_{L,S}}{1 - \Gamma_{L,S}} \right) \cdot Z_0. \tag{2.41}$$

The small-signal amplifier design equations introduced in this section assume a device that is *unconditionally stable*, such that there are no terminations within the unitary circle of the Smith chart (*i.e.* non-negative impedance values) that result in the oscillation of the device. The stability of a device can be assessed through the *Rollet's*

condition, which states that the condition necessary and sufficient for unconditional stability is:

$$\begin{cases} K = & \frac{1-|S_{11}|^2-|S_{22}|^2+|\Delta|^2}{2|S_{12}S_{21}|} > 1 \\ |\Delta| = & |S_{11}S_{22} - S_{12}S_{21}| < 1 \end{cases} \tag{2.42}$$

If a device is not unconditionally stable, it is *potentially unstable*. Note that this *does not mean* that the device is necessarily unstable under any possible termination. From a practical point of view, most microwave transistors are either unconditionally stable or potentially unstable with $K < 1$ and $|\Delta| < 1$. Negative values of K in the range $-1 < K < 0$ result in a transistor instable for most of the terminations within the unitary circle of the Smith chart. In case of a potentially unstable device, the choice of the desired terminations must be done by combining the gain contours with the plot of the device stability region (also analytically derived), which can result in a trade-off between gain and stability. Alternatively, a potentially unstable device can be stabilized by using zero-pole cancellation techniques or resistive dumping, which in turn reduces the gain.

2.5 FET RF Power Amplifiers

The power amplifier represents the last block of the transmitter front-end chain, and its function is to boost the RF signal delivering to the load (*i.e.* the antenna) the maximum power possible without introducing an excessive level of distortion. Under such conditions, the device is pushed to its limits in terms of voltage and current RF output swing. Moreover, the reduced gain available when operating the device at high-frequency makes the design of mm-wave power amplifiers an extremely challenging task for both device and circuit designers.

During AC operations, the transistor output current swing is associated to a voltage swing at the load connected to the device output contacts that, as shown in the left side of Figure 2.10, in turn results in a voltage swing at the device output contacts

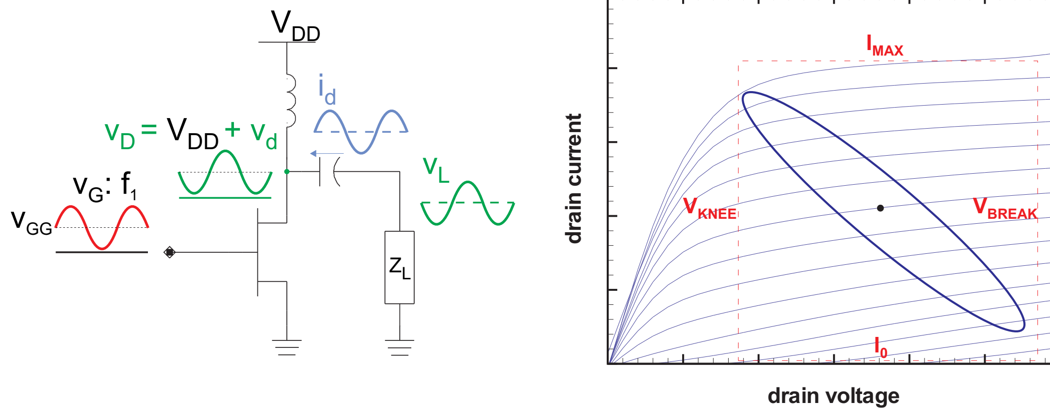

Figure 2.10: Example of a dynamic load-line overlaid to the static $I_D - V_D$. V_{KNEE} is the drain saturation voltage, V_{BREAK} is the breakdown voltage, I_0 is the zero current at the threshold voltage, and I_{MAX} is the device maximum current.

as well. The combination of the output current and voltage swing is called a "*load-line*", and each point of the load-line represents the dynamic position in the output current-voltage space (*IV-space*) of the device during the AC operations. An example of a load-line can be seen in the right of side of Figure 2.10, where the FET drain current swing was represented as of function of the drain voltage swing and overlaid to the static DC $I_D - V_D$ characteristic. This "dynamic" load-line has in general an elliptical shape, as the load connected to the output (which sets the output current and voltage amplitude/phase relation) can generally have a complex value. The device IV output continuously "moves" in the IV-space along the dynamic load-line during the AC operations, and one full loop of the ellipsoid represents one AC swing time-period. All these ellipsoids are perfectly overlaid once the device achieves an AC steady-state situation.

The output drain current swing is driven by the input gate voltage swing through the FET transconductance gain, while the output drain voltage swing is due to the drain current and the load-line. On the other hand, the load also indirectly affects the drain current through the interaction between the drain voltage swing and the device out-

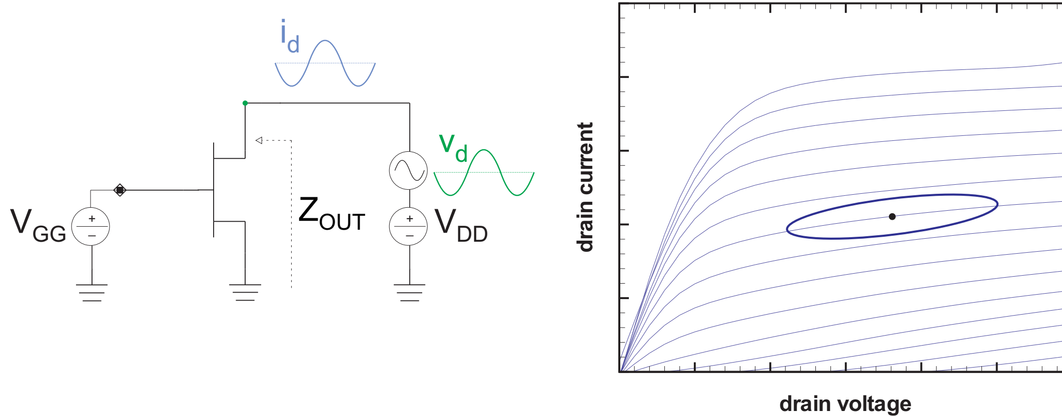

Figure 2.11: Output impedance measurement set-up and resulting drain current swing versus drain voltage swing overlaid to the $I_D - V_D$. The major axis of the ellipsoid is aligned with the slope of the $I_D - V_D$ curves that is associated, in case of small-signals, to the real part of the device output impedance (*i.e.* R_{ds} or more commonly known as r_0, which in FETs is usually due to the Drain Induced Barrier Lowering, *DIBL*).

put impedance, which being in general complex, sets another constraint in terms of current and voltage amplitude/phase relation as shown in Figure 2.11. Therefore, the dynamic load-line incorporates these constraints (the driving gate voltage, the complex load, and the complex FET output impedance) and provides valuable information about the potentials applied to the device contacts, which are directly related the inner carrier distribution and carrier dynamics inside the device during large-signal operations [14]. The gate voltage swing sets to which current curve the load-line is bound in the IV-space at a certain instant, while the drain voltage swing defines the position along this current curve. The two combined ”instantaneous” constraints result in the dynamic load-line. It must noted that in general the transconductance has a complex value, especially at mm-wave frequencies, that takes into account the delay between gate voltage and resulting drain current.

In the left side of Figure 2.10 we can also see that during the large-signal swings, associated to high power applications, the $I_D - V_D$ sets the limits of the dynamic load-line swing that are given by the drain saturation voltage (voltage lower bound to avoid

non-linear operations, also known as the *knee voltage*), breakdown voltage (voltage upper bound to avoid avalanche breakdown), threshold voltage (current lower bound), and device maximum current (current upped bound). However, the *static* DC $I_D - V_D$ does not necessary correspond to the *dynamic* AC $I_D - V_D$ due to the fact that the device dynamically ”moves” in the IV space during the AC operations. For instance, surface traps can lead to a difference between static and dynamic $I_D - V_D$ due to the DC-to-RF current dispersion described in Section 6.1.

The well-known maximum power transfer theorem states that in order to deliver the maximum active power to a load impedance, connected to a source with a finite internal series impedance, the load impedance must be the complex conjugate of the source impedance. This corresponds to impedance values with same real part and opposite imaginary part. This theorem is easily demonstrated with basic circuit theory calculations and can be found in any circuit theory book, but it can also be intuitively understood. The opposite sign of the imaginary parts of source and load impedance cancel out each other, eliminating in such way any reactive part in the overall impedance and the associated reactive power. On the other hand, concerning the delivered active (real) power for a certain real source series impedance, a large real load impedance results in large load voltage but small current, while a small real load results in small load voltage but large current. Thus, it is not surprising that the situation that maximizes the current-voltage product (*i.e.* the power) is the one where the real part of the load equals the real part of the source impedance.

It is interesting to note that this power transfer mechanism, and the interaction between the FET output impedance (*i.e.* the power source with the associated source impedance) and the load impedance connected to the device, is actually embedded in the $I_D - V_D$ characteristics produced as output by the device simulator. This means that the device behavior simulated by the CMC supplies a complete picture that allows one to satisfy the discussed maximum power transfer theorem when the effect of a load-line

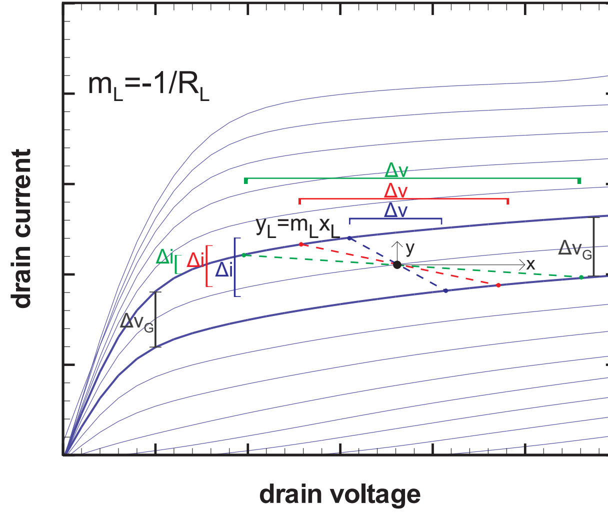

Figure 2.12: Example of different load-lines whose extremes are bound to the highest and lowest current set by the gate voltage swing, Δv_G. The drain current, Δi, and the drain voltage, Δv, swings are also reported for the different cases. The output power is associated to the $\Delta i \cdot \Delta v$ product.

is coupled to the device simulator as will be discussed in Chapter 7 and 8. The proof of this statement, not found in any textbook to the best of the author's knowledge, can be done with some simple geometrical and basic algebra. Referring to Figure 2.12, we can see that the extremes of the real part of the load-lines associated to different load impedances, R_L, are bound to certain maximum and minimum current curves by the maximum voltage swing, Δv_G. On the other hands, the maximum drain voltage swing, due to the load-line, constrains the position of the load-line extremes along these current curves. Keeping in mind that the maximum AC power is associated to the maximum AC current-voltage swing product, $\Delta i \cdot \Delta v$, and that the slope of the load-lines is $m_L = -1/R_L$, we can see that the load-line associated to a large R_L (green case) results in large voltage swing but small current swing. On the other hand, the load-line associated to a small R_L (blue case) results in a large current swing but small voltage swing. Two extreme cases of the just discussed situations are, respectively, an infinite load impedance (*i.e.* max Δv, $\Delta i = 0$) and a zero, short-circuit, load impedance (*i.e.*

max Δi, $\Delta v = 0$). It is evident that the situation (*i.e.* the load-line slope) that maximizes the $\Delta i \cdot \Delta v$ product must lays somewhere in between.

The load-line slope that maximizes the power can be found with some basic algebra and by assuming that with small AC signals all the different current curves have the same constant slope $m_{ds} = 1/R_{ds}$, where R_{ds} is the FET output impedance. The lines associated to the current curves in the proximity of the bias point are described by

$$y_{ds} = m_{ds} \cdot x_{ds} + q_{ds}, \tag{2.43}$$

where q_{ds} value discriminates whether a particular line passes thorough the bias point, above, or below. The load-line can be described as:

$$y_L = m_L x_L, \tag{2.44}$$

where $m_L = -1/R_L$, and y_L and x_L can be associated to, respectively, $\Delta i/2$ and $\Delta v/2$.

Keeping in mind that the load-line extremes are bound to belong to the two extreme current curves, they must also satisfy the current curve Eq.(2.43):

$$\begin{aligned}
y_L &= m_{ds} \cdot x_L + q_{ds} \\
P_L &= y_L \cdot x_L = m_{ds} \cdot x_L^2 + q_{ds} \cdot x_L \\
\frac{dP_L}{dx} &= 2 \cdot m_{ds} \cdot x_L + q_{ds} \\
&\Rightarrow x_{LPmax} = -\frac{q_{ds}}{2 \cdot m_{ds}} && (2.45) \\
&\Rightarrow y_{LPmax} = m_{ds} \left(-\frac{q_{ds}}{2 \cdot m_{ds}} \right) = -\frac{q_{ds}}{2}. && (2.46)
\end{aligned}$$

Now, the slope associated to this maximum power load-line is simply:

$$m_L = \frac{y_L - 0}{x_L - 0} = -\left. \frac{q_{ds}}{2} \middle/ \frac{q_{ds}}{2 \cdot m_{ds}} \right. = -m_{ds} = -\frac{1}{R_{ds}}. \tag{2.47}$$

This results in:

$$R_L = R_{ds}, \;\; QED. \tag{2.48}$$

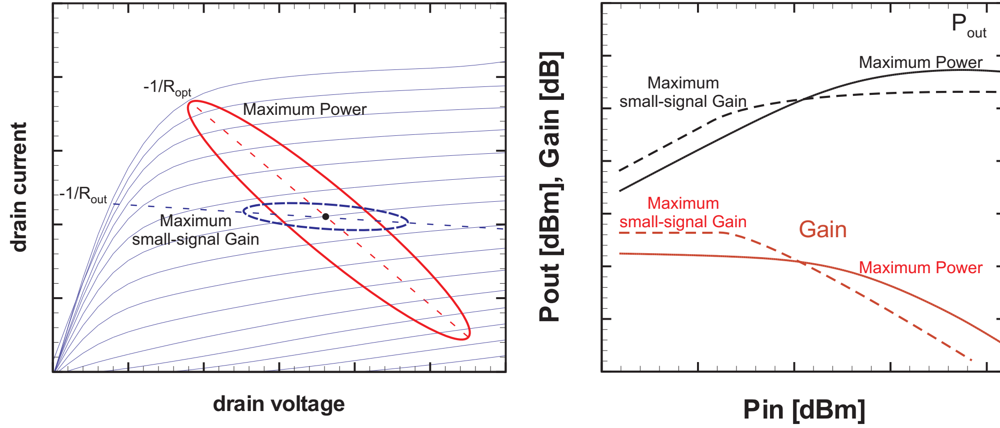

Figure 2.13: Load matching, R_L, for Maximum Small-signal Gain ($R_L = R_{out}$) versus Maximum Large-signal Output Power ($R_L = R_{opt}$), where R_{out} is the real part of the FET output impedance and R_{opt} is the real part of the optimum impedance that maximize the drain and current swing under large-signal operations.

When applied to an amplifier, the maximum power transfer conditions also maximizes the gain for a certain set value of input power. On the other hand, in power amplifiers the large-signal current-voltage swing can be limited by the $I_D - V_D$ boundaries as discussed before. In particular, as shown in from the left side of Figure 2.13, a load impedance satisfying the small-signal maximimum gain condition (*i.e.* $R_L = R_{out} = R_{ds}$) would not allow to obtain full power (*i.e.* full current-voltage swing product), with respect to another optimum load-line (*i.e.* $R_L = R_{opt}$) that maximize the absolute output power, because as the input power is increased the associated large-signal load-line will end up prematurely hitting the $I_D - V_D$ boundaries. Therefore, the load impedance matched for maximum small-signal gain does not in general allow to achieve the maximum output power possible. This issue is shown in the right side of Figure 2.13, where we can see that for small input powers (*i.e.* small-signal condition) the largest power, and the largest gain, is achieved for a load matched for maximum small-signal gain. On the other hand, as soon as the input power is increased, the $I_D - V_D$ limits seriously reduces the performance of the amplifier matched with such

a load. Conversely, the amplifier matched with a load for maximum absolute power, achieves the best power performance for large input powers. In both cases, the output power saturates for large input power, as it will be discussed in Section 2.5, but this saturation point occurs later when the amplifier is matched for maximum absolute power due to the optimized load-line.

Another issue related to large-signal operations is that the small-signal assumption of constant FET output impedance is not justified anymore, since the slope of the different $I_D - V_D$ curves changes dynamically in the IV-space. Moreover, the concept of output impedance, defined as an univocal complex value that sets the voltage current amplitude/phase relation, is hard to define as it can been seen from Figure 8.16. Furthermore, the other small-signal intrinsic parameters also dynamically change during the large-signal device operations as the device internal carrier distribution changes with the large-signal time-varying value of the gate and drain potentials. These issues are discussed the next sections.

Non-linear Effects

Input dependent-parameters are usually associated to non-linear systems, so a FET in large-signal regime exhibits strong non-linearities. As already discussed in the previous paragraph, the dependence of the device parameters from the input is due to the large potential swing at the contacts as well as to the limits set by the dynamic $I_D - V_D$ boundaries, beyond which an input-dependent distortion of the output waveform occurs.

In general, the output response of a non-linear system can be expanded as a Taylor series of the input signal voltage, v_i:

$$v_o = a_0 + a_1 v_i + a_2 v_i^2 + a_3 v_i^2, \tag{2.49}$$

where a_i are the Taylor coefficients. For a single tone input, which is the most common case for continuous wave (CW) power amplifier characterization, the input can be

expressed as:

$$v_i = V_0 \cos \omega_0 t, \tag{2.50}$$

and the resulting output voltage is:

$$\begin{aligned} v_0 =& a_0 + a_1 V_0 \cos \omega_0 t + a_2 V_0^2 \cos^2 \omega_0 t + a_3 V_0^3 \cos^3 \omega_0 t + \ldots \\ =& \left(a_0 + \frac{1}{2} a_2 V_0^2 \right) + \left(a_1 V_0 + \frac{3}{4} a_3 V_0^3 \right) \cos \omega_0 t + \frac{1}{2} a_2 V_0^2 \cos 2\omega_0 t \\ &+ \frac{1}{4} a_3 V_0^3 \cos 3\omega_0 t + \ldots \end{aligned} \tag{2.51}$$

As we can see, the output response contains other frequency components, generated by the non-linear behavior, beyond the fundamental frequency. In particular, these generated frequency components correspond to the higher harmonics (*i.e.* integer multiples) of the single-tone fundamental frequency in input.

The situation is more complicated when we have multiple input tones, which is the case in almost all the real applications when a non-pure sinusoidal waveform is fed as input signal to the device. In particular, for the simplest case with two input tones:

$$v_i = V_0(\cos \omega_1 t + \cos \omega_2 t), \tag{2.52}$$

the resulting output voltage is:

$$\begin{aligned} v_0 =& a_0 + a_1 V_0 (\cos \omega_1 t + \cos \omega_2 t) + a_2 V_0^2 (\cos \omega_1 t + t \cos \omega_2 t)^2 \\ &+ a_3 V_0^3 (\cos \omega_1 t + \cos \omega_2 t)^3 + \ldots \\ =& a_0 + a_1 V_0 \cos \omega_1 t + a_1 V_0 \cos \omega_2 t + \frac{1}{2} a_2 V_0^2 (1 + \cos 2\omega_1 t) + \frac{1}{2} a_2 V_0^2 (1 + \cos 2\omega_2 t) \\ &+ a_2 V_0^2 \cos(\omega_1 - \omega_2) t + a_2 V_0^2 \cos(\omega_1 + \omega_2) t \\ &+ a_3 V_0^3 \left(\frac{3}{4} \cos \omega_1 t + \frac{1}{4} \cos 3\omega_1 t \right) + a_3 V_0^3 \left(\frac{3}{4} \cos \omega_2 t + \frac{1}{4} \cos 3\omega_2 t \right) \\ &+ a_3 V_0^3 \left[\frac{3}{2} \cos \omega_2 t + \frac{3}{4} \cos(2\omega_1 - \omega_2) t + \frac{3}{4} \cos(2\omega_1 t + \omega_2 t) \right] \\ &+ a_3 V_0^3 \left[\frac{3}{2} \cos \omega_1 t + \frac{3}{4} \cos(2\omega_2 - \omega_1) t + \frac{3}{4} \cos(2\omega_2 + \omega_1) t \right] + \ldots \end{aligned} \tag{2.53}$$

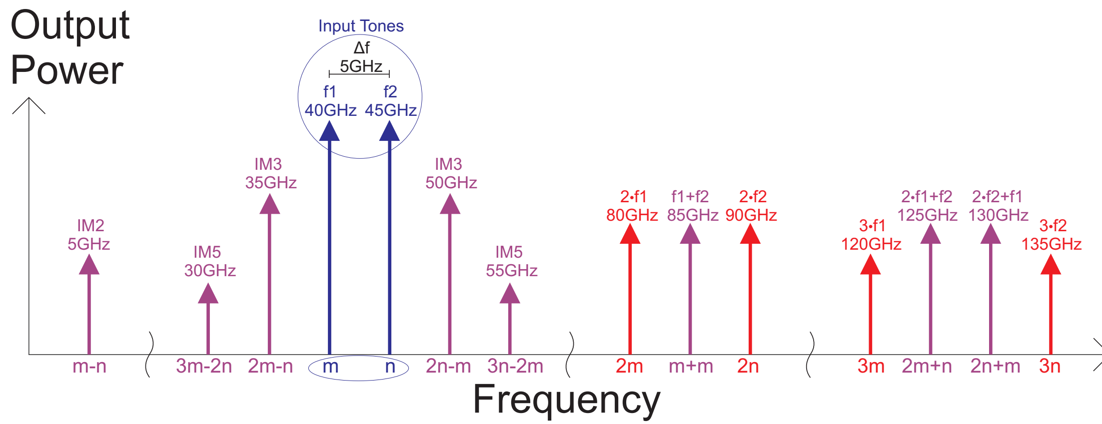

Figure 2.14: Output spectrum of a non-linear system, stimulated by a two-tones input, showing the higher harmonics and the intermodulation product spurious frequency components up to the 3rd order.

As we can see, the output spectrum consists of several components, shown in Figure 2.14, whose frequency can be expressed in the following general form:

$$m\omega_1 + n\omega_2, \tag{2.54}$$

where m and n are integer numbers, and the the *order* of each generated frequency component is given by $m+n$. In particular, the frequency components associated to both m and n different from zero are called *intermodulation products*. All these spurious frequencies, harmonics and intermodulation products, are undesired in the amplifier output and can be in general filtered-out by a selectively enough pass-band filter. In addition, the very high-frequency ones will be also cut-off by the device intrinsic output capacitance. However, we can see that the third-order intermodulation products, $IM3$, corresponding to $(2\omega_1 - \omega_2)$ and $(2\omega_2 - \omega_1)$, are located near the original input tones, so they are the most problematic because they can be within the band of the filter.

Another main difference between linear and non-linear system is the validity of the superposition principle. As we can see from Figure 2.15, the superposition principle does not hold in non-linear systems due to the frequency generation and the interaction between input tones. In particular, the output response of a linear system is, in the worst case, an attenuated/amplified and delayed version of the original input tone. In the case

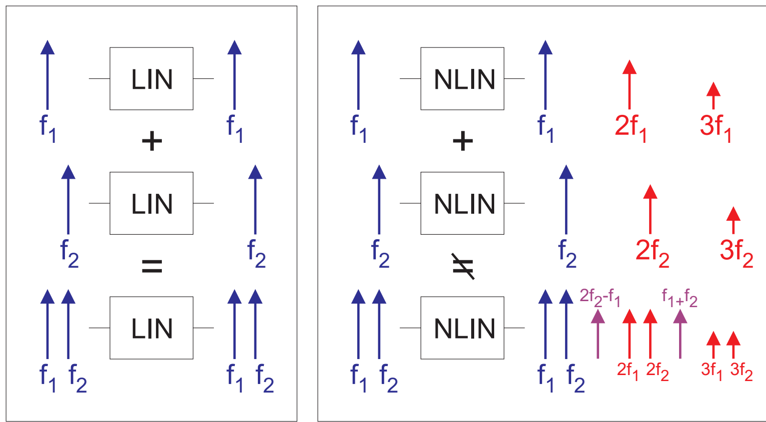

Figure 2.15: Example of superposition of input tones in a linear (left) and non-linear system (right).

of multiple input tones, resulting in output tones with different amplitude and phase, this corresponds to a *linear distortion*. In any case, the response of a linear system is the sum of the response to the single input tones, which is not the case for non-linear systems.

The discussed non-linear effects are just the most relevant for the purpose of this work. Memory effects, often due to self-heating and surface traps (described in Section 6.1) introduce additional sources of non-linearity in mm-wave *GaN* HEMT power amplifiers. A more complete discussion about non-linear systems and non-linear microwave circuits can be found in the Maas book [61].

Large-signal versus Small-signal Analysis

The limit of the small-signal approach lays in the assumptions that the amplitude of the AC signal is much smaller than the DC biasing, so that the system exhibits a linear response. However, non-linearity arises as soon as the amplitude of the AC component of the signal increases and makes the higher order terms of the Taylor expansion reported in Eq.(2.49) not negligible. In such situation, the signal cannot be considered small anymore. Therefore, under large-signal conditions, the small-signal analysis is no longer appropriate due to its linearity assumption and the unequivocal values assigned

to the intrinsic parameters during the device operations. Such parameters are changing with time in large-signal regime according to the large swing voltage at the input and output of the device. Even though our Monte Carlo device simulator can accurately reproduce the internal physics of a device under any type of semi-classical regime, what is lacking in the case of the large-signal regime is an adequate and standard approach that associates the circuit model parameters with the simulated device. In this case, many of the theoretical performance predictions (*i.e.* maximum gain, optimum load impedance, gain contours), based on the small-signal analysis models described in Section 2.4 cannot be considered valid.

Considering an FET specifically, the small-signal characterization provides a description through Y-, Z-, S-parameters, input and output impedance, gain, and small-signal intrinsic parameters for each frequency. In a linear-system, there is not interaction between frequency components, thus the measurement of parameters for a specific frequency value is independent of the other harmonic components of the signal. However, non-linear effects, such as harmonic generation and intermodulation, lead to strong interactions between different frequency components. Furthermore, the input-dependence determines different small-signal parameters at different input power levels. The dominant non-linear parameters for a FET are C_{gs}, C_{gd}, R_{ds}, and g_m. For example, there is a strong dependence of the transconductance (g_m) on the gate voltage (v_G) (typically evident in a DC $I_D - V_G$ curve, see for example Figure 3.5), which affects the device large-signal gain depending on the input power. A variable gain in large-signal operations implies a distorted output signal in relation to gain compression and harmonic generation, as will discussed in the next sections. In addition, small-signal analysis cannot take into account the role of the whole $I_D - V_D$ output characteristics in limiting the maximum output current and voltage swing. Moreover, the fact that the dynamic $I_D - V_D$ differs from the DC one makes harder to characterize and predict the actual device performance during normal large-signal and high-frequency operations.

Thus, the device behavior in the large-signal regime is the effect of a complex situation where a standard effective parameterization has still to be devised, and where design equations are much more complicated (and may require more draconian approximations) than the small-signal ones reported in Section 2.4. So, the prediction of device performance through a parameterized model in a fashion similar to the small-signal framework is not simply feasible. Specific to power amplifier applications, it is difficult to define the meaning of output impedance in the large-signal regime, and the impact of the dynamic load-line on gain and maximum output power makes it difficult to determine the optimal load impedance. Within the small-signal framework, the optimal load for both maximum gain and maximum output power is represented by the complex conjugate of the output impedance. However, for large-signal operations, the issues just discussed do not allow the analytic prediction of the optimal load value in an exact way. Experimentally, the *load-pull* technique [62, 54] is extremely popular for large-signal characterization. Within this approach, the load value is swept within the device stability range (which is commonly assessed through small-signal analysis), and the optimal load is empirically found with an extended number of measurements once the large-signal output power contours are plotted. For each tested load, the source impedance is adjusted to guarantee the maximum power transfer.

An example of load-pull technique results, obtained through through the Agilent Design System (ADS) microwave circuit simulator, is reported in Figure 2.16. As we can see, for low input power levels ($P_{in} = -10$ and 0 dBm), the power contours have a circular shape that follow the analytical equations discussed in Section 2.4. On the other hand, as soon as we increase the power close to the $1dB$ gain compression point ($P_{in} = 10$ dBm) of the amplifier, as will be described in the next section, the contours do not resemble circles anymore. Furthermore, the largest measured input power ($P_{in} = 20$ dBm) results in output contours far from the analytical description of the circular shape as well as we can verify that the load that maximizes the power (*i.e.* the

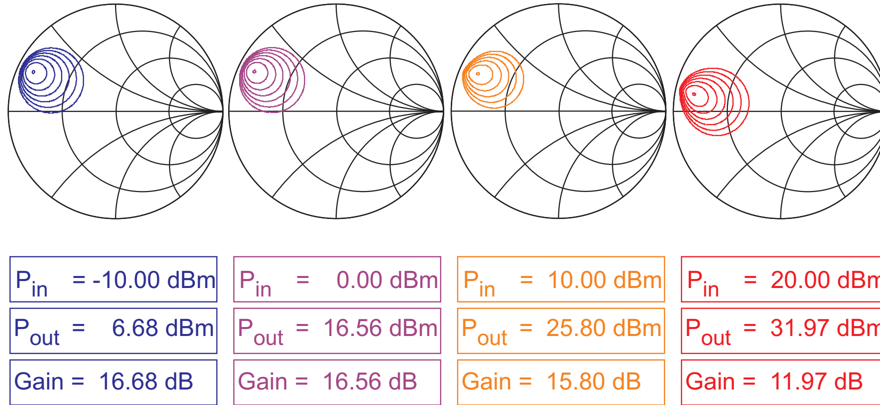

Figure 2.16: Example of constant output power contours, P_{out}, plot in the Smith chart for different input power levels, P_{in}. The contours were obtained by using the load-pull technique applied to a FET amplifier simulated through the Agilent Design System (ADS) microwave circuit simulator. The FET compact model was part of the ADS tool. The inner most point corresponds to the load that maximizes the output power. For low input power levels, the contours follow the analytically prediction of the small-signal analysis in terms of circular shape, load that maximizes the power corresponding to the conjugate matching, and center of the circles laying along the line connecting the optimum impedance to the center of the Smith chart.

most point inner point) is shifted with respect to the one associated to low input power (*i.e.* small-signal analysis conjugate matched load). Moreover, the maximum absolute output power is achieved at the expenses of the amplifier gain (see Section 2.5 and the next one). Similar empirical output power contours can be used to determine the optimum source impedance (*source-pull*).

Power Amplifier Figures of Merit

The most important figures of merit for a power amplifier are the *output power*, P_{out}, and the *gain*, G. In particular, as the power amplifier is in general a non-linear system, these figures of merit are also dependent on the input power and are usually plotted as function of the input power, P_{in}.

An example of a typical single-tone CW input power sweep characterization is reported in figure Figure 2.17 for a Class-A power amplifier (described in Section 2.5). As we can see, P_{out} linearly increases with P_{in} in the low power range. The non-linear

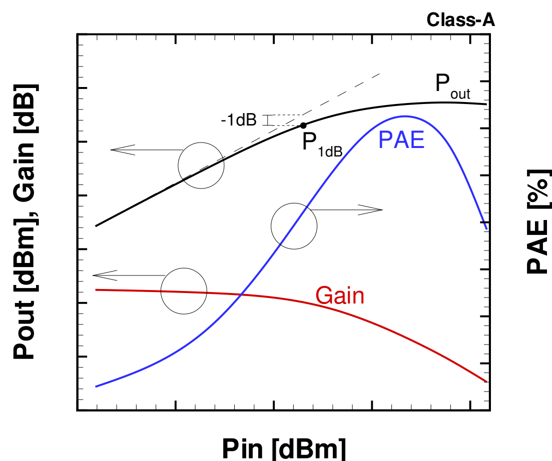

Figure 2.17: Example of typical single-tone input large signal figures of merit: Output power (P_{out}), gain, and power added efficiency (PAE) vs input power, P_{in} at the fundamental frequency. The 1dB compression point, P_{1dB}, is also reported.

distortion at higher power results in a reduction of the slope of P_{out} at the fundamental frequency as P_{in} is increased (*i.e.* the gain drops). In particular, the input power is transferred to the higher harmonics, rather than to the fundamental frequency, due to the amplifier non-linear behavior. This phenomenon is known as *gain compression* or *output power saturation*, and it is in general mainly associated to the increment of the Taylor series coefficient a_3 (negative for amplifiers), which results in the generation of a third harmonic about 180° out of phase with the output wave form as well as the reduction of the fundamental frequency amplitude as evident in Eq.(2.51).

The *1dB compression point*, P_{1dB}, is defined as the point where, for a ginve P_{in}, P_{out}, or conversely the gain, is 1 dB below the P_{out} linear value. The 1dB compression point can be expresses either in terms of output power (*output referred 1dB compression point*) or in terms of input power (*input referred 1dB compression point*). Commonly, the 1dB compression point represents the operating point where a good trade-off between output power and linearity can be achieved. The low input power region where P_{out} linearly increases with P_{in} is associated to a constant gain, which is known as *linear gain*, G_{lin}, and corresponds to the small-signal gain.

Another key figure of merit for power amplifiers is the *power added efficiency*, PAE, which represents the efficiency evaluated through the incremental RF power added by the amplifier versus the cost in terms of DC power consumption, P_{outDC}:

$$PAE = \frac{P_{out} - P_{in}}{P_{outDC}} \cdot 100\% = \frac{P_{out} - P_{in}}{I_{Ddc} \cdot V_{Ddc}} \cdot 100\%. \tag{2.55}$$

In general, the higher is V_D the larger can be the drain voltage RF swing that does not drive the amplifier in the non-linear region below the knee (saturation) voltage, avoiding in such way the reduction of the output power at the fundamental frequency as discussed in the previous sections. It must be noted that P_{outDC} is not necessarily constant for increasing P_{in}; this is because an asymmetric load-line (*i.e.* due distortion or asymmetrical operation like the one in Class-B amplifiers described in Section 2.5) can result in a DC drain current and voltage "dynamic" DC components different from the DC bias ones. As we can see from Figure 2.17, PAE is low for low input power due to the fact V_D is much larger than the value required to keep the drain voltage RF swing above the knee voltage. On the other hand, the PAE achieves its maximum for a specific value of P_{in} ,that allows a good trade-off between DC consumption and available swing for current and voltage.

We can also define the *drain efficiency*, DE (also indicated by η):

$$DE = \eta = \frac{P_{out}}{P_{outDC}} \cdot 100\% = \frac{P_{out}}{I_{Ddc} \cdot V_{Ddc}} \cdot 100\%, \tag{2.56}$$

which is also particularly common for the characterization of RF oscillator sources, where the oscillation is initiated without the need of RF input power.

While the *1dB compression point* characterizes the performance of the device in terms of power and linearity, a more complete figure of merit related to the non-linearity of the device is the *third-order input referred intercept point*, $IIP3$, obtained by measuring one of the third-order intermodulation products, $(2\omega_1 - \omega_2)$ and $(2\omega_2 - \omega_1)$, resulting from a two-tone input. As we can see from Figure 2.18, this intercept point is obtained as the intersection between the extension of the linearly increasing output

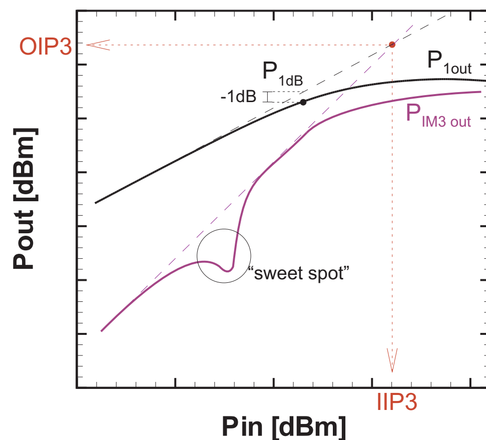

Figure 2.18: Example of typical two-tone input large signal figures of merit: Fundamental frequency output power (P_{1out}) and third-order intermodulation frequency output power, P_{IM3out}, vs input power, P_{in}. The third-order input referred intercept point, $IIP3$, and the third-order output referred intercept point, $OIP3$, are also reported.

power of the fundamental frequency, P_{1out}, and the extension of the linearly increasing output power, P_{IM3out}, of one of the two intermodulation frequencies. Moreover, we can also see that in general P_{IM3out} does not necessarily increase monotonically with increasing power, but the curves can exhibit a dip that has been sometimes observed in HEMTs and *GaAs* MESFETs, and is known as the *sweet spot*, and occurs under some particular conditions of intermodulation phase cancellation [63, 64].

RF Power Amplifier Classes

In this section, a brief description is offered of the most relevant amplifier classes for high-frequency operations. More details can be found in [61] and [62]. The different classes are defined by the device bias scheme, which heavily affects the output waveform shape, the amplifier efficiency, and the gain.

As discussed in the previous sections, high-power amplifiers often operate under non-linear conditions resulting in the generation of undesired out-of-band frequency components such as higher harmonics and intermodulation products. So, a pass-band output filter is commonly included in the amplifier design in order to suppress these

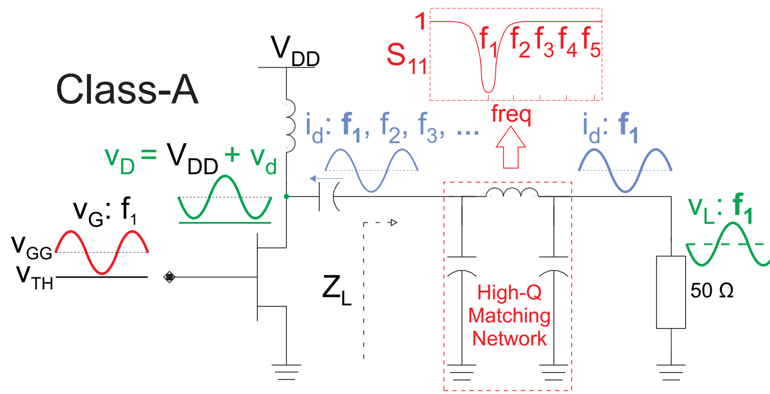

Figure 2.19: Typical Class-A RF power amplifier output matching configuration.

undesired spurious spectral components. In practical applications, the output matching network can also be used for this purpose as it can be designed by using large reactive elements in order to have a high Q factor, achieving in such way a high selectivity tuned around the fundamental frequency. A narrow band set up is commonly used for single-tone and two-tone high-power characterization of the transistor regardless whether the actual application will be wide band or narrow band amplifiers.

CLASS-A

A typical Class-A amplifier set up is shown in Figure 2.19, while the drain current and voltage waveforms are shown in the left side of Figure 2.20. As we can see, this amplifier class is biased with a gate bias corresponding to a drain current which is roughly half of the maximum drain current of the device. This allows the maximum current swing possible without distortion, and results in a good degree of linearity as long as the output power does not exceed the $I_D - V_D$ boundaries.

Thus, as long as the amplifier is used within its linear region such that the generation of spurious frequency components is limited, the constraints on the high-Q matching network can be relaxed allowing wide band applications as well. On the other hand, a Class-A amplifier always consumes DC power even when there is no RF input, and its drain efficiency by definition cannot exceed 50% since the DC drain current, I_{Ddc}, is always half the peak amplitude of the drain current swing, I_{peak}.

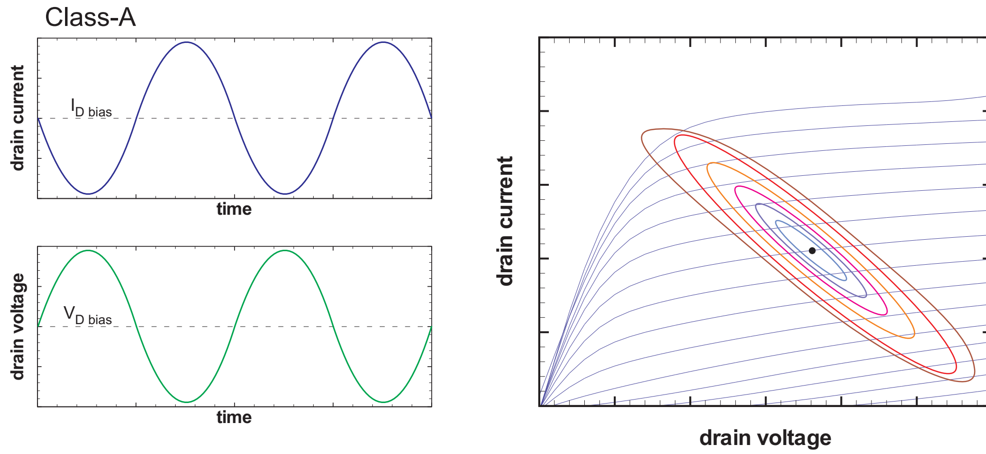

Figure 2.20: Example of Class-A output current and voltage waveforms (left) and dynamic load-line overlaid to the static $I_D - V_D$ (right). Different dynamic load-lines for different increasing input power levels are also reported.

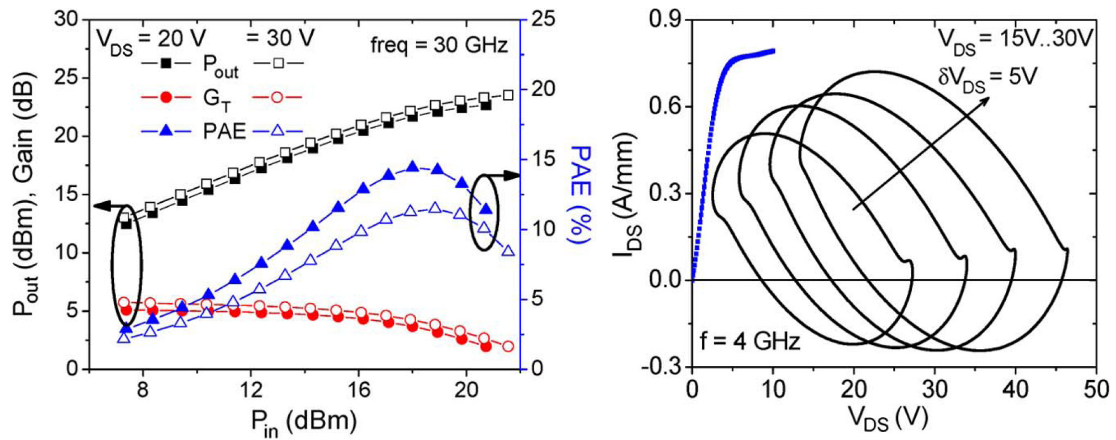

Figure 2.21: Example of experimental large-signal chracterization of a 100-μm-wide *GaN* HEMT biased near Class-A operations [4]. The 30 GHz power characterization is shown in the left panel, while the 4 GHz dynamic load-lines are shown in the right panel.

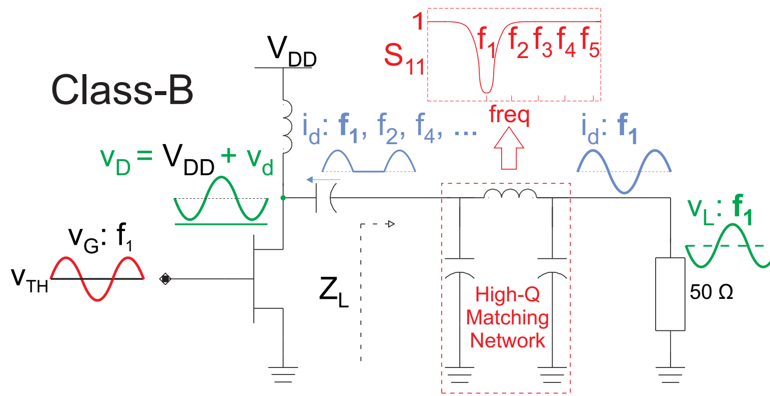

Figure 2.22: Typical Class-B RF power amplifier output matching configuration.

CLASS-B

A typical Class-B amplifier set up is shown in Figure 2.22 As we can see, this amplifier class is biased with a gate bias corresponding to the threshold voltage such that the DC power consumption is zero when there is no RF input signal. On the other hand, this results in an half-sinusoidal output drain voltage waveform due to the fact that the transistor is not conducting for half of the input cycle. Therefore, the drain current spectrum inherently has a large component of higher harmonics due to the highly non-linear on/off switching operation. The second harmonic is particularly strong due to the even half-sinusoidal waveform, and it can be only 7.5 *dB* below the fundamental one if the device is terminated with a load that has the same impedance value at the fundamental and at the second-harmonic frequency [61]. However, a device biased for Class-B can still act as linear amplifier, in sense of input-output power proportionality, as long as a high-Q matching network is used to filter off the higher harmonics and recover the sinusoidal current waveform at the fundamental. This sinusoidal current is subsequently delivered to the load resulting in a sinusoidal voltage load-line as shown in the left side of Figure 2.23. This is because the resonant circuit, composing the matching network, is charged during the conducting portion of the device cycle and discharges into the load when the device is off providing the sinusoidal outputs desired [5]. Therefore, a

high-Q matching network is mandatory in case of Class-B amplifier device characterization, unless other more complex topologies are used such as the common push-pull configuration described in the next section.

The asymmetric output current waveform results in the dependence of the "dynamic" DC point of the device on the output current waveform amplitude, I_{peak}, which is in turn due to the input power. The DC value of the drain current, I_{Ddc}, is equal to I_{peak}/π in a half sinusoidal waveform, and therefore the Class-B amplifier DC point changes with the input power (self-biasing effect). This also results in a linear gain that slightly increases for increasing low input power due to the fact that, as I_{Ddc} increases from zero, the device gets closer to the biasing point corresponding to the peak transconductance as shown in the right side of Figure 2.23

The theoretical maximum efficiency of a Class-B amplifier is 78.5% [61], which is larger than the Class-A one due to the fact that I_{Ddc} is always lower than half of I_{peak}. On the other hand, a Class-B amplifier has an inherently lower gain than a Class-A one because a FET biased at the threshold voltage needs an input gate voltage swing which is twice the one needed for a Class-A biasing to achieve the same I_{peak}. Moreover, such large voltage swing can reduce the input-output power linear proportionality for the reasons discussed in the previous sections.

CLASS-AB

A circuit topology, commonly used in the design of the actual amplifier rather than during the device characterization, is the *push-pull* configurations that avoids the need for a high-Q output filter in Class-B amplifiers. In such topology, the full sinusoid is recovered by using two complementary devices (*i.e.* NMOS and PMOS technologies) such that by combining the two generated half-sinusoids a full sinusoid is delivered to the load.

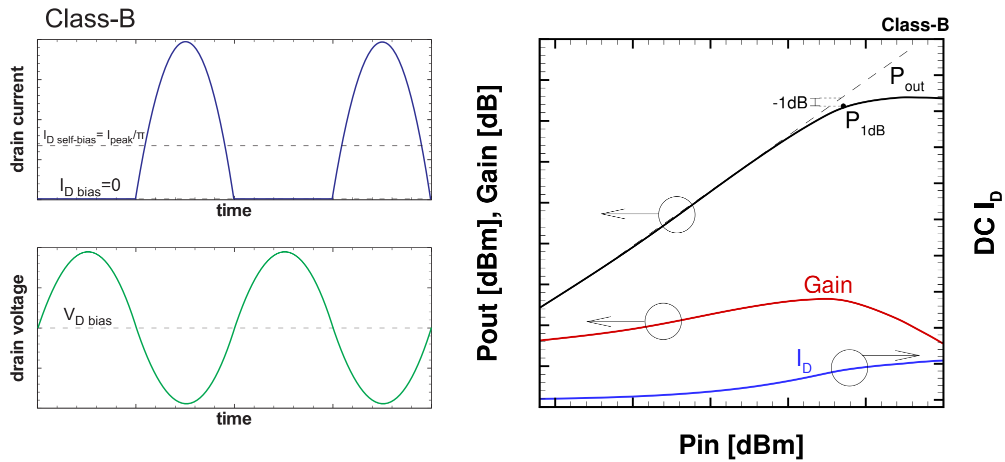

Figure 2.23: Example of Class-B output current and voltage waveforms at the device output contacts (left) and large-signal characterization (right): Output power (P_{out}), gain, and DC drain current, I_D, vs input power, P_{in}.

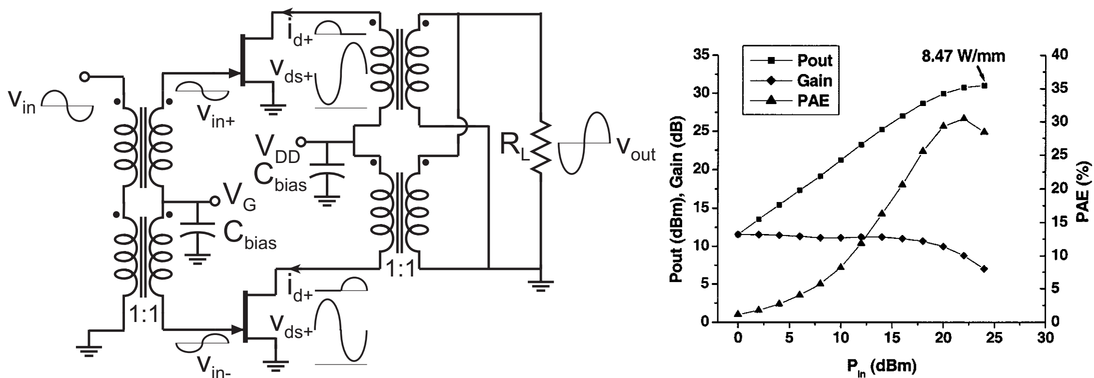

Figure 2.24: LEFT: Example of a typical push-pull amplifiers topology implemented by using non-complementary FETs biased for deep Class-AB operations [5]. RIGHT: Example of experimental large-signal power performance of a 150-μm-wide $AlGaN/GaN$ HEMT biased for Class-AB operations [6] at 8 GHz.

The push-pull configuration can be also implemented in transistor technologies where complementary devices are not available (which is the case for HEMTs) by using the circuit topology reported in Figure 2.24. In such non-complementary configuration, the excitation is applied 180° out-of-phase to the inputs of the two transistors biased for Class-B operations, and the output waveforms are subsequently combined 180° out-of-phase. This effectively reconstructs a full sinusoidal output signal by suppressing the even harmonics, the only ones ideally generated by a Class-B amplifier, associated to a half sinusoidal waveform. In such way, wide band operations are also possible by the relaxation on the Q factor of the output matching network. In any case, each transistor produces a half-sinusoidal drain current waveform and has a sinusoidal drain voltage waveform in both the single transistor Class-B configuration and the push-pull dual transistor configuration.

However, the issue with real FETs biased for Class-B operations is that the soft pinch-off of the $I_D - V_G$ characteristic does not allow an ideal on/off behavior resulting in a problematic recovery of the pure sinusoidal waveform at the output of the push-pull amplifier (introducing the so called *cross-over distortion* where the two out-of-phase half sinusoids are joint). Moreover, the fact that the transconductance is zero at the turning on point (*i.e.* the biasing point) results in a very small small-signal gain (see Figure 3.5 for an example). This also implies a large variation of the gain for a variable input power, which represents a problem for communication schemes using a multilevel modulation where waveforms with a non-constant envelope result in a large Peak to Average Power Ratio [65]. For these reasons, the so called Deep Class-AB is particular popular because the FET is biased slightly above the threshold voltage, commonly about 10% of I_{peak}, to avoid such issues while still preserving a small enough DC power consumption. Experimental device characterization is commonly performed on a single transistor biased for deep Class-AB operation, even though the actual circuit design will likely be a push-pull configuration with two transistors.

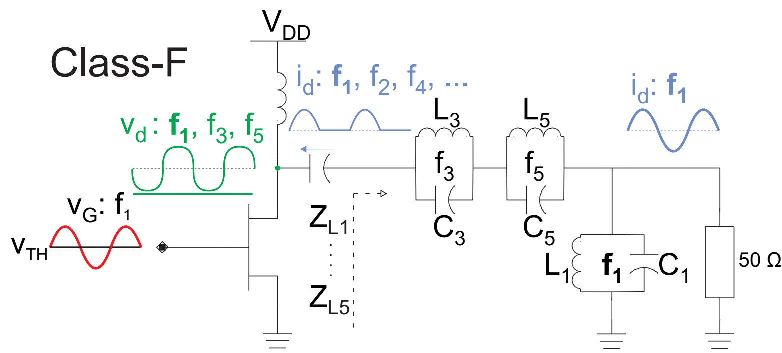

Figure 2.25: Typical Class-F RF power amplifier output matching configuration with the parallel resonant tanks tuned at the first, third, and fifth harmonic resulting in a drain voltage waveform with a content of only odd harmonics.

CLASS-F

A Class-F amplifier is a special variant of the Class-B one where not all the higher harmonics are short-circuited, such that different optimized impedance values are associated to the different harmonic frequencies (*i.e. harmonic loading*). An early paper by Snider [66] identified optimum terminations for transistor power amplifiers, and reported that the optimum terminations are short circuits at even harmonics and open circuits at odd harmonics other than the fundamental.

An example of a typical Class-F amplifier with harmonic loading for the first five harmonics is shown in Figure 2.25. In such a configuration, the half-sinusoidal output current waveform of the transistor biased for Class-B ideally presents only even harmonics. On the other hand, the output matching network, characteristic of Class-F, is composed by parallel resonant LC tanks. An ideal LC tank (infinite Q factor) shows an open circuit at the resonant frequency, and a short circuit at any other frequency. As we can see in Figure 2.25, there are parallel LC tanks (L_3C_3 and L_5C_5) connected in series to the load and tuned at the higher odd harmonics (*i.e.* third and fifth). Such tanks therefore result in an ideally infinite impedance termination at the resonant frequency that forces the presence of significant odd harmonics in the drain voltage waveform

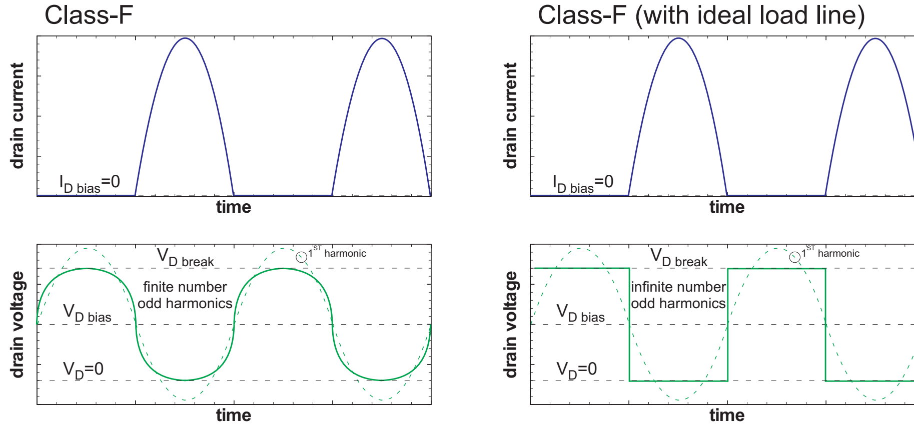

Figure 2.26: Example of Class-F output current waveform and output voltage waveform, with a finite number of odd harmonics, at the device output contacts (left), and current waveform and output voltage waveform, with a infinite number of odd harmonics (right).

even though the drain current waveform odd harmonics are weak or vanishingly small. On the other hand, the even harmonics, which are strong in the drain current waveform, are suppressed in the voltage waveform due to the fact that all the LC tanks (tuned at the fundamental, third, and fifth harmonic: shunt parallel L_1C_1 tank, and series parallel L_3C_3 and L_5C_5 tanks) ideally act as short circuit, effectively shorting the even harmonics by creating a zero-resistance conductive path to ground. At the fundamental frequency, the LC tanks tuned at the odd harmonics (L_3C_3, and L_5C_5) do not interfere (*i.e.* series short circuits), while the one tuned at the fundamental frequency (L_1C_1) acts as impedance matching network by matching the device output to the load.

All of these ideal Class-F operations are based on the assumption that the transistor can be modeled as an ideal current source (*i.e.* independent of the drain voltage, resulting in flat $I_D - V_D$ due to an infinite device output impedance), such that the output voltage can be shaped to any value independently of the current flowing. In such a way, odd harmonics can be forced in the drain voltage waveform without inducing odd harmonics in the drain current waveform as well. As we can see from the left

m, n Denotes Maximum Order of Harmonics in Drain Current and Voltage, Respectively	$n = 1$	$n = 3$	$n = 5$	$n = \infty$
$m = 1$	50%	57.7%	60.3%	63.7%
$m = 2$	70.7%	81.7%	85.3%	90.0%
$m = 4$	75%	86.6%	90.5%	95.5%
$m = \infty$	78.5%	90.7%	94.8%	100%

Figure 2.27: Maximum efficiency of Class-F Power Amplifiers [7, 8]. The $m = 1$ and $n = 1$ case corresponds to a Class-A amplifier, while the $m = \infty$ and $n = 1$ case corresponds to an ideal Class-B case. The $m = \infty$ and $n = \infty$ case represents an ideal Class-F amplifier with square current and voltage waveforms.

side of Figure 2.26, the introduction of voltage odd harmonics, in particular the third harmonic 180° out-of-phase, results in a compression of the voltage waveform that allows to increase the amplitude of the compressed drain voltage swing closer to the $I_D - V_D$ boundaries while taking the fundamental frequency amplitude beyond these limits without incurring in distortion. This results in an effective larger fundamental frequency swing at the load (*i.e.* larger power delivered to the load). Furthermore, the theoretical large efficiency achievable by Class-F amplifiers can be seen in the right side of Figure 2.26, where the voltage waveform is composed by an infinite number of odd harmonics (practically forced by using a $\lambda/4$ transmission line instead of the series connected parallel LC tanks [7]). In such cases, since the overlap between the current and voltage waveforms at the device contacts is zero, the instantaneous power dissipated by the transistor is zero. On the other hand, both the current and voltage swing at the load are fully sinusoidal due to the Class-F output network. Therefore, the theoretical efficiency can be as high as 100% if the current even harmonics are also infinite (*i.e.* square current waveform). On the other hand, when a finite number of harmonics is used, there is a finite dissipation in the transistor due to the partial overlap between current and voltage. Pratically, up to five harmonics are usually included, with a theoretically maximum efficiency of 90% as shown in Figure 2.27

Class-F amplifiers achieve better efficiency than Class-B ones because the output harmonics are reactively tuned. On the other hand, the actual non-infinite and complex output impedance, as well as the non-ideal non-exclusively even Class-B waveform, results in the fact that the odd harmonics will not necessarily be optimally tuned with an infinite impedance, and the even harmonics with an impedance equal to zero. In actual Class-F amplifiers, the optimal impedance for odd harmonics will be generally large (and close to a phase shift of 180° in the third harmonic case), while the optimal impedance for the even harmonics will be generally small and often including an imaginary part to cancel out the device output intrinsic capacitance [13].

Chapter 3

Emerging N-face GaN HEMT Technology

3.1 Introduction

In contrast to cubic III-V semiconductors like *GaAs* and *InP* with the zincblende structure, the thermodynamically stable phase of *InN*, *GaN* and *AlN*, is the hexagonal wurtzite structure [41]. Due to the inversion asymmetry along the *c* axis in the wurtzite phase, structures grown along the (0001) and $(000\bar{1})$ directions, giving rise to *Ga*- and *N*-polarity respectively [67], have different surface properties and growth kinetics [68]. All the polarization directions are flipped over in *Ga*-face material, as compared to *N*-face materials [1]. This enables *N*-face heterojunction device designs where the 2DEG forms *above* the wide-bandgap layer which then acts as an inherent back-barrier as shown in Figure 3.1.

Ga-face devices have largely been investigated in the past few years, mainly because of the more challenging fabrication process of device-grade *N*-face films. However, new high quality *N*-face nitride processes were recently developed [68, 69], making the *N*-face reverse polarization very attractive for device design. *N*-face HEMTs, indeed, show several potential advantages over *Ga*-face devices [24, 23]. In particular, the wider-bandgap *AlGaN* layer, which is located below the *GaN* channel layer in *N*-face devices, provides a natural back-barrier to the channel electrons when the transistor is biased near pinch-off, improving the confinement of the carriers, particularly in deep submicrometer devices [70]. At the same time, the smaller bandgap of the top *N*-face *GaN* layer allows for easier access to the two dimensional electron gas (2DEG) channel than through the *AlGaN* barrier layer like the *Ga*-face case, making possible the fabrication of non-alloyed ohmic contacts with low access resistance [71].

In *Ga*-face $AlGaN/GaN$ structures, the activation yield of *Si*-implanted dopants in the top *AlGaN* barrier is extremely low, which increases the difficulty of fabricating ultra-low contact resistances [12]. On the other hand, in *N*-face samples the top layer

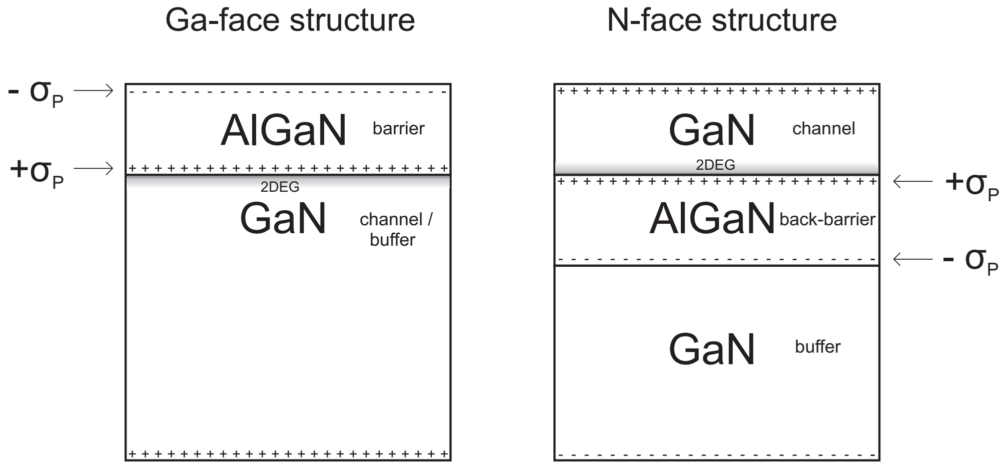

Figure 3.1: Comparison between typical Ga-face and N-face HEMT vertical structures.

is the *GaN* channel where the activation yield is very high due to the *AlGaN*-free surface [71]. Furthermore, a low *Al* composition *AlGaN* cap can be employed to reduce the gate leakage, due to the lower *GaN* Schottky gate barrier, to values even lower than in *Ga*-face devices [68]. The low contact resistance can be preserved by etching the *AlGaN* cap in the ohmic regions. Moreover, it has been shown that *N*-face *InGaN* can be easily grown, comparable to *In*-polar, with high-quality properties [72], leading to a promising exploitation on high-velocity *InGaN* channel *GaN* HEMTs. *Mg* doping can be obtained without the problems of surface polarity inversion on the *N*-polar surface [73]. Lastly, the *N*-face *GaN* channel can be depleted by the negative polarization charge of an *AlGaN* cap grown on its top in the gate region [74]. Thus, *N*-face devices can be easily fabricated for enhancement mode operation without using a gate recess [75, 76], avoiding the difficult threshold voltage control related to such technique.

Recently, 150 nm gate length *N*-face metal-insulator-semiconductor (MIS) HEMTs with a current gain cut-off frequency f_T of about 47 GHz and a maximum oscillation frequency f_{max} of 81 GHz have been reported [10, 4]. Such performance has been made possible by novel design techniques, including the use of an *Al* barrier in order to eliminate alloy scattering while providing strong charge confinement [73]. High perfor-

mance *N*-face *GaN*/*AlGaN* devices grown on a sapphire substrate using metal-organic chemical vapor deposition are described in [77]. A new method of fabricating *N*-face $GaN/AlGaN$ HEMTs through layer transfer technology was introduced by Jinwook *et al.* [70]. This fabrication technique allows for *N*-face devices with a maximum drain current 30% higher than in *Ga*-face devices, and an RF performance comparable to the state-of-the-art *Ga*-face devices.

In this chapter, we first introduce the theory of spontaneous and piezoelectric polarization in Wurzite group-III-nitrides. Then, we demonstrate the agreement between our simulation approach and experiment, focusing in particular on a device layout proposed by Nidhi *et al.* [10]. We subsequently analyze the effects of the gate length scaling on both the DC and RF device performance. We then investigate the effects of shortening the access regions, and finally, we study the enhancement configuration of the *N*-face structure.

An overview of the polarization characteristics of HEMT devices is presented in Section 3.2. The structure of the device analyzed is described in Section 3.3, where the simulation setup is summarized as well. The effects of the gate scaling are discussed in Section 3.4. In Section 3.5, the effect of shortening the access regions is studied. The performance of enhancement mode devices is investigated in the next section. Conclusions are finally drawn in Section 3.7.

3.2 Polarization Engineering in GaN HEMTs

Wurzite group III-nitrides show large polarization fields that can be exploited for device engineering purposes. The spontaneous and piezoelectric polarization of *AlN* and *GaN* can indeed be up to ten times larger than in conventional III-V (*i.e.* *GaAs*) and II-VI cubic compounds [78]. Polarization discontinuities in $AlGaN/GaN$ heterostructures result in a sheet of net positive polarization-charge at the $AlGaN/GaN$ heterointerface. Free electrons will tend to compensate this polarization charge, creating a two dimensional electron gas which acts as the channel in *GaN* HEMTs. In particular,

Table 3.1: Ideal and predicted real *c/a* ratio comparison for binary wurtzite group III nitrides [1]

	InN	GaN	AlN
a	3.585 Å	3.199 Å	3.110 Å
ideal *c/a*	1.633 Å	1.633 Å	1.633 Å
real *c/a*	1.618 Å	1.634 Å	1.606 Å

the field induced by the piezoelectric polarization, due to strain at the $AlGaN/GaN$ heterostructure, is more than five times larger than in $AlGaAs/GaAs$ heterostructures, with values up to 2 MV/cm. The spontaneous polarization can cause electric fields up to 3 MV/cm [9]. Furthermore, the spontaneous polarization increases with the non-ideality of the crystal structure, and is found to be larger in *AlN* than in *GaN* due to the larger mismatch between ideal and real (obtained from experiments as well as theoretical predictions) c/a ratio in *AlN* layers [1] (see Table 3.1). Thus, *AlGaN* layers show larger polarization with increasing *Al* concentrations. This allows for high carrier concentration at the heterointerface even for relaxed heterostructures (*i.e.* in absence of piezoelectric field), because of the difference in spontaneous polarization between *GaN* and *AlGaN* materials.

Strain in epitaxial layers is due to lattice mismatch and/or mismatch of the thermal expansion coefficients with the substrate layer. This strain leads to a piezoelectric effect along the $[0001]$ direction parallel to the c-axis in wurtzite *GaN* based heterostructures [79]. The resulting piezoelectric polarization can be obtained from the polarization coefficients, which are almost one order of magnitude larger than in conventional III-V materials, such as *GaAs*, due to the stronger polar nature of the group-III nitrides [80].

Assuming that the length of the hexagonal edge of the relaxed unit cell is a_0, and that the unit-cell height is c_0, the piezoelectric polarization induced along the c axis

can be written as [1]:

$$P_{PE} = e_{33}\varepsilon_z + e_{31}(\varepsilon_x + \varepsilon_y), \tag{3.1}$$

where the coefficients

$$\varepsilon_z = \frac{c - c_0}{c_0}, \varepsilon_x = \varepsilon_y = \frac{a - a_0}{a_0}, \tag{3.2}$$

are the strains along the c axis and in the basal plane, respectively. Here, e_{33} and e_{31} are the piezoelectric coefficients, while a and c are the lattice constants of the strained layer. The relationship between lattice constants in $AlGaN$ is given by

$$\frac{c - c_0}{c_0} = -2\frac{C_{13}}{C_{33}}\frac{a - a_0}{a_0}, \tag{3.3}$$

where C_{13} and C_{33} are the elastic constants.

The piezoelectric polarization is given by substituting (3.2) and (3.3) into (3.1):

$$P_{PE} = 2\frac{a - a_0}{a_0}\left(e_{31} - e_{33}\frac{C_{13}}{C_{33}}\right), \tag{3.4}$$

where a_0 is the length of the hexagonal edge of the relaxed unit cell, a and c are the lattice constants of the strained layer, e_{33} and e_{31} are the piezoelectric coefficients, and C_{13} and C_{33} are the elastic constants. The orientation of the polarization is defined such that the positive direction is from the N anion towards the Ga or Al cation. Furthermore, since the inequality

$$\left(e_{31} - e_{33}\frac{C_{13}}{C_{33}}\right) < 0, \tag{3.5}$$

is valid for all Al compositions in the $AlGaN$ system, the piezoelectric polarization is negative for tensile strain and positive for compressive strain [81]. Therefore, the polarization vector points from the cation to the anion for tensile strain, and *vice versa* for compressive strain.

In epitaxial heterostructures, when a thin $AlGaN$ film is grown on a thick GaN layer, the strain in the $AlGaN$ layer is tensile, whereas the resulting strain is compressive when a GaN layer is grown on a relaxed $AlGaN$ layer. Since the spontaneous

polarization values of group III - nitrides is negative [81], the orientations of the piezoelectric and spontaneous polarization vectors in the $AlGaN$ layer are parallel in the case of tensile strain, and anti-parallel in the case of compressive strain. The total polarization within a layer can then be expressed as

$$P = P_{PE} + P_{SP}, \tag{3.6}$$

where P_{PE} and P_{SP} are the piezoelectric and spontaneous polarization, respectively. The polarization changes substantially within two heterostructure layers, and the total net polarization is where P_{PE} and P_{SP} are the piezoelectric and spontaneous polarization, respectively. For the heterostructure being considered here, the polarization changes substantially within the bilayer, and the total net interface polarization-induced charge is associated with the gradient of the polarization

$$\sigma_P = -\bigtriangledown P. \tag{3.7}$$

Let us consider now a heterostructure system with a thin $Al_xGa_{1-x}N$ layer grown on an $Al_yGa_{1-y}N$ layer, having a different molar fraction. If $y < x$, the $AlGaN$ layer would be under tensile strain, and the spontaneous and piezoelectric polarization vectors would point in the same direction. If $y > x$, the strain is compressive and the resulting polarization vectors are anti-parallel.

Linear interpolation of the parameters representing the physical properties of AlN and GaN can be used [9] in order to calculate the polarization-induced charge density in the alloys. The spontaneous polarizations in the two layers, $P_{SP}(x)$ and $P_{SP}(y)$, can then be expressed as

$$\begin{aligned} P_{SP}(x) &= [xP_{SP_{AlN}} + (1-x)P_{SP_{GaN}}], \\ P_{SP}(y) &= [yP_{SP_{AlN}} + (1-y)P_{SP_{GaN}}]. \end{aligned} \tag{3.8}$$

As a result, the polarization-induced sheet charge density at the pseudomorphically-grown $Al_xGa_{1-x}N/Al_yGa_{1-y}N$ heterointerface yields:

$$
\begin{aligned}
\sigma_P &= \left[P_{Al_yGa_{1-y}N} - P_{Al_xGa_{1-x}N}\right]\hat{\mathbf{u}}_p\cdot\hat{\mathbf{x}} \\
&= \{[P_{SP}(y) - P_{SP}(x)] - P_{PZ}(x)\}\,\hat{\mathbf{u}}_p\cdot\hat{\mathbf{x}} \\
&= [P_{SP_{AlN}} - P_{SP_{GaN}}](y-x)\hat{\mathbf{u}}_p\cdot\hat{\mathbf{x}} \\
&\quad - 2\frac{a(y)-a(x)}{a(x)}\left(e_{31(x)} - e_{33}(x)\frac{C_{13}(x)}{C_{33}(x)}\right)\hat{\mathbf{u}}_p\cdot\hat{\mathbf{x}},
\end{aligned}
\tag{3.9}
$$

where

$$
\begin{aligned}
a(x) &= [xa_{AlN} + (1-x)a_{GaN}], \\
a(y) &= [ya_{AlN} + (1-y)c_{GaN}],
\end{aligned}
\tag{3.10}
$$

are the lattice constants of the epitaxial and the substrate layers, respectively. $\hat{\mathbf{u}}_p$ is the unit vector that points from the *N* anion towards the *Ga* or *Al* cation, while $\hat{\mathbf{x}}$ is the unit vector that defines the growth direction. The piezoelectric coefficients of the $Al_xGa_{1-x}N$, $e_{33}(x)$ and $e_{31}(x)$, as well as the elastic constants $C_{33}(x)$ and $C_{31}(x)$, can also be obtained by linear interpolation.

The different growth directions and atomic layer terminations for *Ga*-face and *N*-face materials define the direction of positive polarization as shown in Figure 3.2. In *Ga*-face terminated layers, the *Ga* anions are on the top surface, so the positive polarization direction points toward the surface. Thus, the piezoelectric polarization due to the spontaneous and the tensile strain points toward the substrate, while the compressive piezoelectric polarization points toward the surface for *Ga*-face. The positive direction points toward the substrate for *N*-face terminated layers where the *N*-cations

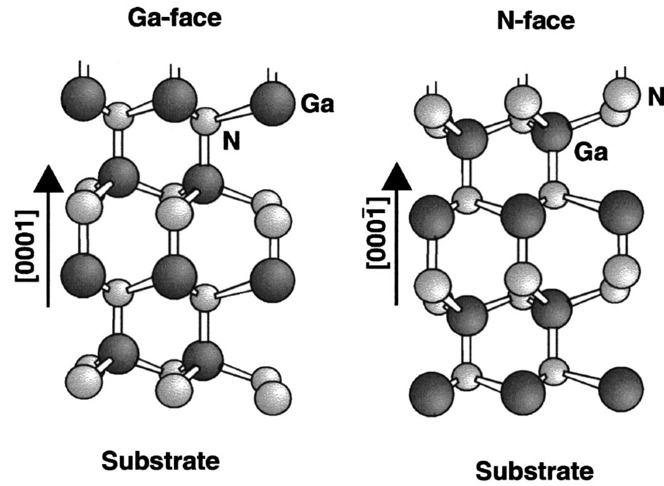

Figure 3.2: Schematic drawing of the crystal structure of wurtzite Ga-face and N-face *GaN* [9].

are on the top surface. Hence, all the polarization directions are flipped over in *Ga*-face devices, as compared to *N*-face ones.

Therefore, different combinations of polarization discontinuities in *AlN*/*GaN* (or *AlGaN*/*GaN*) heterostuctures can result in a sheet of positive polarization net charge at the *AlN*/*GaN* interface, as shown in Figure 3.3. For instance, tensile strained *AlGaN* grown over relaxed *Ga*-face *GaN* (Figure 3.3.b) will result in an electron channel located at the top side of the *Ga*-face *GaN* lower layer. For the *N*-face case, a sheet of positive polarization net charge at the heterointerface is obtained for compressively strained *N*-face *GaN* grown over relaxed *AlGaN* (Figure 3.3.e). This results in a electron channel located at the bottom side of the *N*-face *GaN* upper layer. Similarly to this last case, a *N*-face vertical structure made of, from bottom to top, a relaxed *GaN* buffer, a tensile strained *AlGaN* (*i.e.* thinner than the critical thickness) with lattice constant constraint to be equal to the one of the *GaN* the substrate, and a consequently non-strained *GaN* channel layer, will results in a 2DEG located at the bottom side of this last top layer. This last configuration is the most common for *N*-face *GaN* HEMTs.

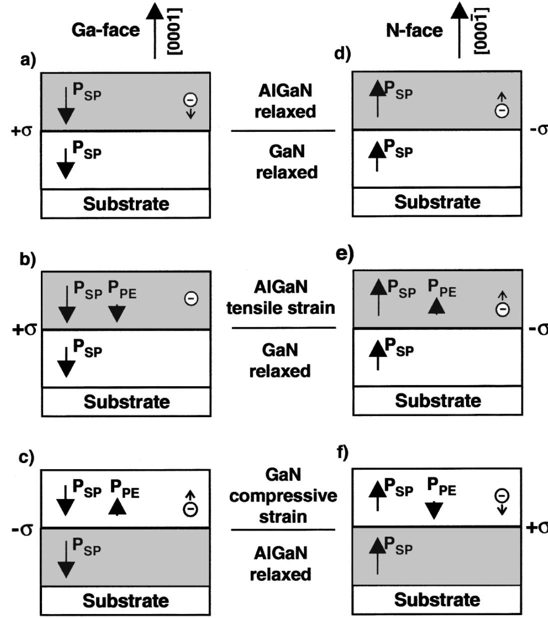

Figure 3.3: Polarization induced sheet charge density and directions of the spontaneous and piezoelectric polarization in Ga- and N-face strained and relaxed *AlGaN*/*GaN* heterostructures [9].

3.3 Device Structure and Simulation Method

The layout of the device described in this work is illustrated in the left side of Figure 3.4, and is made of a *GaN*/*AlN*/*GaN* heterostructure grown on a semi-insulating *SiC* substrate by metal-organic chemical vapor deposition. The heterostructure consists, from bottom to top, of 600 *nm* of unintentionally doped *GaN* followed by a 5.5 *nm Si*-doped *GaN* layer and a 4 *nm GaN* spacer layer. On the top of this stack, the *AlN* barrier and the *GaN* channel layers have been grown. The use of an *AlN* barrier layer instead of *AlGaN* reduces random alloy scattering[1]as a mechanism limiting low-field electron mobility [73] and improves charge confinement [10]. A silicon-doped layer was used to supply charges to the channel and prevent modulation of slow-responding trap states near the valence-band edge at the bottom *AlN*/*GaN* interface [70]. The electrons were further separated from the dopants by a *GaN* spacer layer to reduce remote ionized

[1]While only the fraction of the wave function penetrating the *AlGaN* would feel the random alloy, the presence of the alloy leads to a rougher interface. The latter would, in turn, lead to enhanced scattering of the carriers in *GaN*. We do not distinguish these differences here.

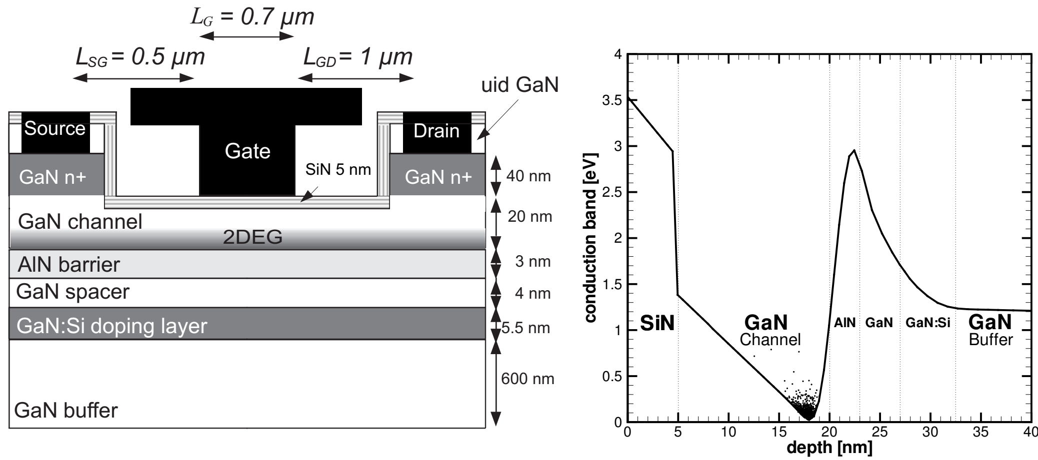

Figure 3.4: LEFT: Schematic cross-section of the simulated *N*-face HEMT. RIGHT: Conduction band profile, and carrier distribution (black points) corresponding to a bias of 0V on the gate and 0V on the drain.

impurity scattering. The source-gate and gate-drain separations are 0.5 μm and 1 μm, respectively, while the gate length is set to 0.7 μm.

As mentioned in the Introduction of this chapter, the presence of the *AlN* layer enhances the confinement of the electrons in the channel. In fact, the large polarization-induced electric field in the *AlN* layer raises the conduction band edge of the *GaN* buffer with respect to the *GaN* channel, creating a potential barrier against carrier diffusion in the bulk layer. This property is highlighted in the right side of Figure 3.4, showing the carrier distribution (black points, at $V_{GS} = 0\ V$, $V_{DS} = 10\ V$) superimposed on the conduction band profile of the simulated device. More details on that device can be found in [10], where all the characteristics of the sample used to obtain the experimental data are reported.

To account for the large polarization discontinuity present at the AlN/GaN heterointerfaces, explicit sheets of charge were placed at these interfaces following the formulation reported in Section 3.2. The large polarization-induced charge of opposite sign present at the two AlN/GaN heterointerfaces compensate each other, due to the

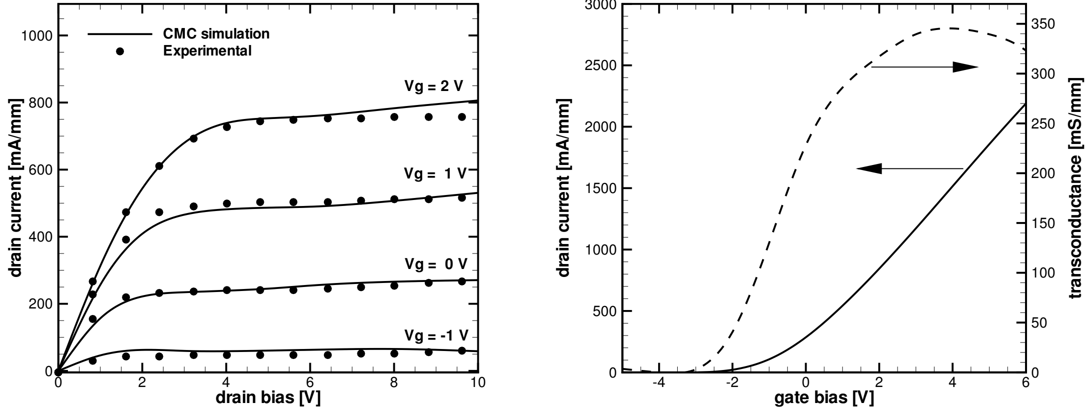

Figure 3.5: LEFT: Comparison of simulated (lines) and measured (dots) $I_D - V_D$ characteristics of the HEMT. The gate voltage V_G is varied between 2 and $-1\ V$ with a bias step of $\Delta V_G = 1\ V$. RIGHT: I_D-V_G characteristic and the corresponding transconductance resulting from the CMC results for a drain bias of $V_D = 10\ V$.

lack of free holes to compensate the negative sheet charge [82]. A silicon-doped layer was therefore used to supply charges to the channel [70]. The material characteristics (e.g. bandstructure or phonon dispersion) are calculated in the basis formed by the primitive vectors of the reciprocal Bravais lattice. The rotation of the crystal needed to simulate N-face GaN is therefore performed by simply rotating these three primitive vectors. Confinement of the 2-DEG at the $AlGaN/GaN$ interface leads to quantum phenomena, such as charge setback and quantization of energy levels [48], which must be taken into account when charge distribution inside a HEMT device is modeled.

In the left side of Figure 3.5, we show a comparison of the simulated (dashed lines) and the measured (dots) I_D-V_D characteristics of the HEMT [10]. As it can be seen, the simulation results show a very good agreement with the experimental data. The I_D-V_G characteristic and the corresponding transconductance are illustrated in the right side of Figure 3.5. The device exhibits a high transconductance peak of about 340 mS/mm for a gate voltage of $V_G = 4\ V$. However, this value is certainly overestimated, since self-heating effects have not been taken into account in the present calculations [11]. Moreover, we should underline that for gate voltages larger than 3 V

we observed the formation of a secondary channel at the top of the SiN/GaN interface. This secondary channel becomes comparable, in terms of electron density, to the main channel for a gate voltage of 6 V. Furthermore, at such high gate voltages, the leakage current due to tunneling through the *SiN* insulator would not be negligible anymore. However, the gate leakage was not investigated in this work.

3.4 Gate Length Scaling

High-power operation has been achieved by *GaN* HEMTs in the millimeter wave frequency range. In order to improve the high-frequency performance, the gate length L_G of the device has to be reduced. However, for short gate lengths, short channel effects such as threshold-voltage shift, soft pinchoff, and high sub-threshold current occurs, due to the poor confinement of the electrons in the 2DEG channel, which reduces the modulation efficiency of the gate [83]. In the *N*-face structure considered here, the *AlN* layer forms a back barrier reducing short channel effects. The use of *AlN* provides strong back-barrier confinement of the 2DEG under high electric fields for device scaling [73].

Gate Length Scaling Effect on DC Performance

In order to investigate the effects of the gate length scaling on the DC performance, the $I_D - V_D$ characteristics for $L_G = 700\ nm$, $600\ nm$, $450\ nm$, $300\ nm$, $150\ nm$, $100\ nm$, and $50\ nm$ were simulated as shown in Figure 3.6. As it can be seen, decreasing the gate length,increases the drain current – and consequently the transconductance peak value – leading to an improvement of the device performance. However, a threshold shift toward negative values, caused by the drain induced barrier lowering (DIBL) effect, is present for a gate length of 100nm. It is interesting to notice how the peak electric field along the channel decreases with the gate length. As shown in Figure 3.7, the reduction of the channel portion modulated by the gate, and therefore its resistance, leads to a decrease of the voltage drop in this region and, as a consequence, of the electric field at the drain end of the gate.

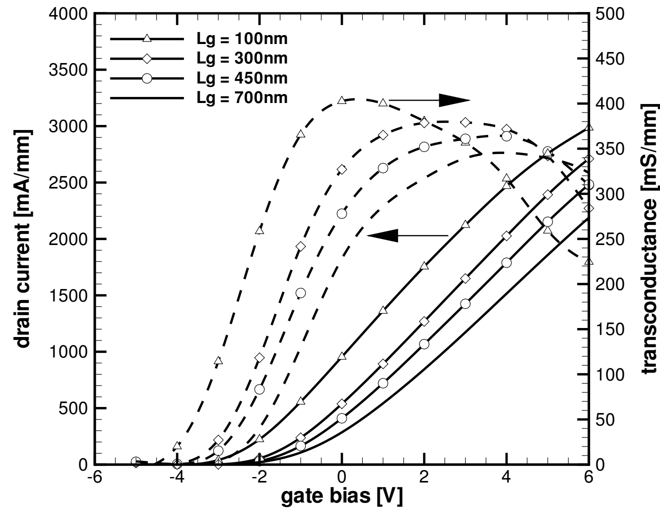

Figure 3.6: $I_D - V_D$ characteristics (solid lines), and transconductance (dashed lines) of the *GaN* HEMT analyzed obtained varying the gate length from L_G = 700 *nm* to L_G = 100 *nm*.

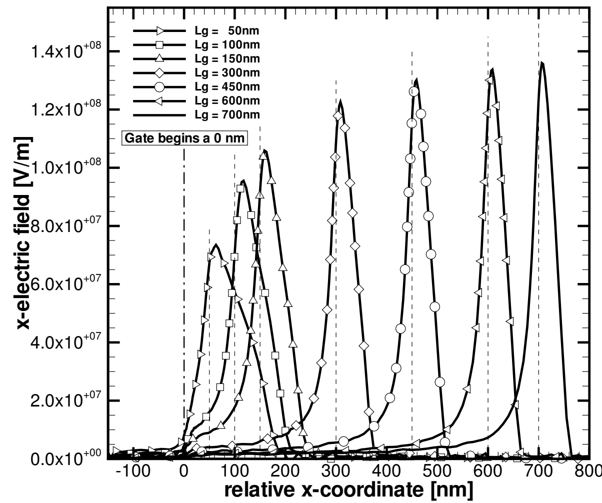

Figure 3.7: Electric field along the channel obtained by CMC simulations performed for $V_D = 10\ V$ and $V_G = 0\ V$, with gate length from $L_G = 700\ nm$ to $L_G = 50\ nm$. The gate source-end (vertical dash-dot line) and the gate drain-end (vertical dashed line) are also plotted for convenience.

To understand how the gate scaling influences the RF performance of the device, we analyzed its frequency response by varying the gate length. The cut-off frequencies for different gate lengths were extracted through small-signal analysis and by evaluating the electron transit time under the gate, as described in Chapter 2, and the results are shown in Figure 3.8. As it can be seen, the cut-off frequency scales almost linearly with the inverse of the gate length down to $L_G = 150\ nm$. This is due to the fact that the scaling reduces the transit time of the carriers across the region under the gate, but does not increase the velocity that is already at the maximum value due to carrier valley transfer, as we can see from Figure 3.9. Moreover, the carriers in the gate region are accelerated up to the maximum velocity overshoot value, but they do not have the time to slow down to the lower maximum steady-state velocity value, which is 2.5×10^5 m/s in GaN. For gate lengths below 150 nm, the decreasing aspect ratio and increasing effective gate length prevent a cut-off frequency linear scaling. This issue related to short-channel effects is further investigated in Chapter 5. The effects of scaling on the electron velocity across the N-face HEMT channel are shown in Figure 3.9 for $L_G =$ 700, 600, 450, 300, and 150 nm. Good agreement between the two f_T extraction methods was found.

3.5 Impact of Access Region on Device Performance

In order to get insight about how the access region lengths influence the performance of the device, both DC and RF analysis were performed on the HEMT device by scaling both the gate-source and gate-drain access region by an equal factor. The resulting DC characteristic are shown in Figure 3.10 where we can see that the decrease of the access region resistances consequent to the adopted scaling leads to an important increase in the device transconductance. Differently from the previous case, decreasing the gate-source and gate-drain distance does not cause short-channel effects and does not change

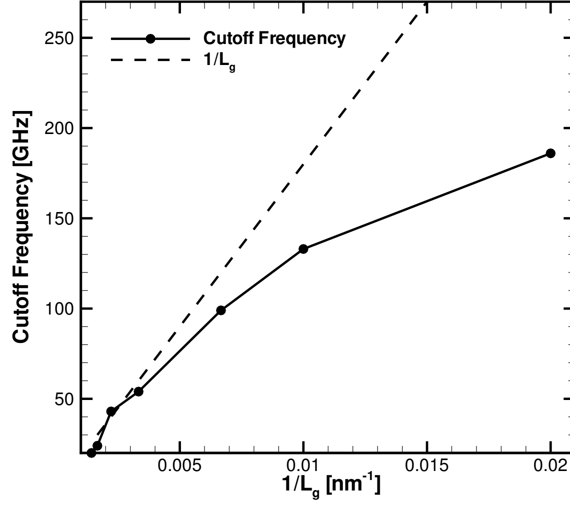

Figure 3.8: Cut-off frequencies of the N-face *GaN* HEMT extracted by the CMC simulation peformed for $V_D = 10\ V$ and a gate biased corresponding to the peak transconductance, with gate length from $L_G = 700\ nm$ to $L_G = 50\ nm$.

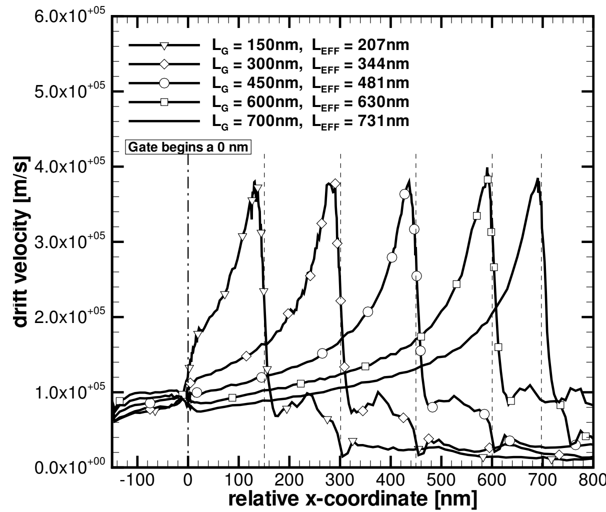

Figure 3.9: Electron drift velocity along the channel obtained by CMC simulations performed for $V_D = 10\ V$ and V_G corresponding to the peak transconductance, with gate length from $L_G = 700\ nm$ to $L_G = 150\ nm$. The gate source-end (vertical dash-dot line) and the gate drain-end (vertical dashed line) are also plotted for convenience.

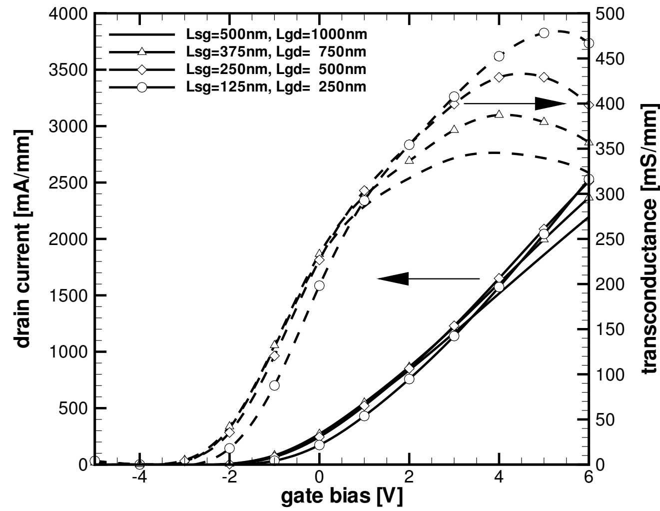

Figure 3.10: $I_D - V_G$ characteristics (solid line) and respective transconductance (dashed line) of the GaN HEMT analyzed obtained for $V_D = 10\ V$, scaling by an equal factor the lengths of both access regions.

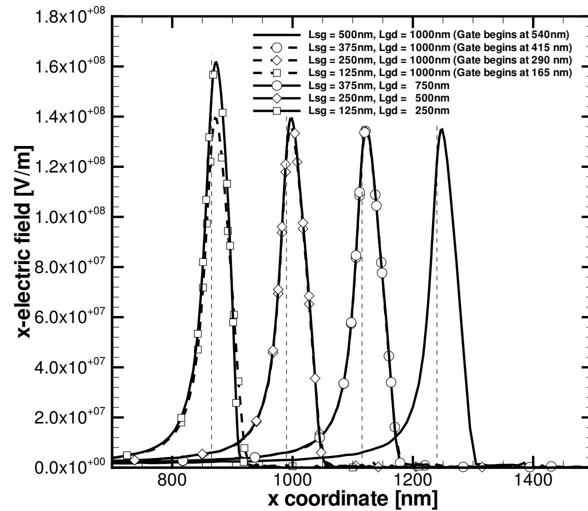

Figure 3.11: Electric field along the channel ($L_G = 700nm$) obtained by CMC simulations performed for $V_D = 10\ V$ and $V_G = 0\ V$, scaling by an equal factor the lengths of both access regions. The the gate drain-end (vertical dashed line) is also plotted for convenience.

the threshold voltage of the device. This peculiarity is an important advantage for circuit design, since it allows a better control of the device characteristics. Figure 3.11 shows how the electric field along the channel changes when varying only the source-gate distance (dashed curves) or the length of both access regions (solid curves). As it can be seen, the size of the source access region does not particularly influence the peak value of the electric field. On the other hand, shrinking the gate-drain distance increases the voltage drop in the region under the gate and, therefore, the maximum electric field in the device. In this case, the frequency response of the device was analyzed. An improvement of the cut-off frequency of about 10% was found for the $L_G = 700$ *nm* device, with access region lengths 75% shorter than the original ones.

3.6 Enhancement and Depletion mode

In the *Ga*-face layouts, an enhancement mode device is created by recessing the *AlGaN* barrier layer in order to obtain a very small gate-to-channel distance. In the *N*-face configuration, the same results can be achieved by adding an *AlGaN* cap layer under the gate insulator (or replacing it with such layer). The reverse polarization field in an *AlGaN* cap grown on a *N*-polar *GaN* channel depletes the channel, and raises the conduction band. This can be used to build an enhancement-mode device without a gate recess [84]. Here, the enhancement mode device has been obtained from the depletion configuration by substituting the *SiN* layer under the gate with an $Al_xGa_{1-x}N$ cap layer. The *AlGaN* cap polarization effect is shown in the left side of Figure 3.12, where the conduction band of the depletion and enhancement mode device are compared for an *AlGaN* molar fraction of $x = 0.5$. For molar fractions larger than $x = 0.5$, and related larger band raising underneath the gate, the effect of the valence band should be taken into account including holes in the simulation

The I_D-V_G characteristics and transconductance of the enhancement *N*-face HEMT for different *Al* concentrations up to $x = 50$ are shown in the right side of Figure 3.12. As expected, by increasing the *Al* molar fraction of the cap layer, the polarization-

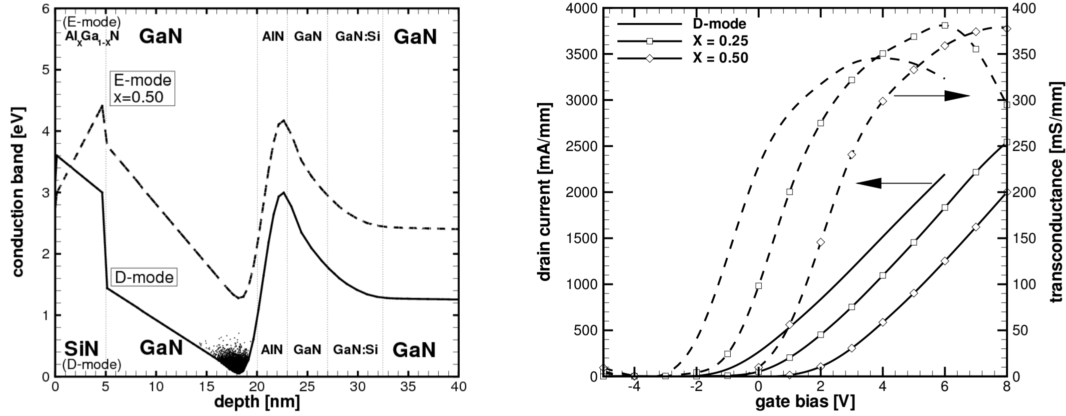

Figure 3.12: LEFT: Conduction band profile for the Depletion (solid line) and Enhancement mode (dashed line) configurations, corresponding to a bias of 0 V on the gate and 0V on the drain. RIGHT: I_D-V_G (solid line) and g_m-V_G (dashed line) characteristic of the enhancement mode N-Face device for different $AlGaN$ molar fraction values evaluated at a drain bias of 10 V.

induced sheet of charge at the upper $GaN/AlGaN$ heterointerface increases, and the threshold voltage consequently shifts toward positive values. The gate leakage current due to the forward biased $Gate/AlGaN$ Schottky junction was not investigated.

3.7 Conclusions

In this chapter, a complete characterization of a state-of-the-art N-face GaN HEMTs been performed using experimental data to calibrate the few adjustable parameters of the simulator. The RF properties of the device have been analyzed through the Fourier decomposition method and related to the electron dynamics in the channel in terms of the average velocity under the gate for different gate lengths. The influence of the access regions on both DC and RF performance has been studied as well. The simulations indicate that shortening the gate-source and gate-drain distances allows a significant improvement in the DC device performance. However, an increase of the maximum electric field is generated by this operation and must be taken into account during the device design phase. Finally, the enhancement mode configuration of the N-face structure has been analyzed.

Chapter 4

Comparison of N-face and Ga-face GaN HEMTs

4.1 Introduction

In spite of their very high potential, *N*-face *GaN* HEMTs show performance that is still inferior to the one of *Ga*-face devices mainly due to the relatively inferior material quality. As the material growth issues of *N*-face structures improve, it becomes of interest to understand whether the *N*-face HEMT technology can achieve significantly better performance than the *Ga*-face technology, and methods to compare the different structures in a systematic way are necessary [23, 24]. In this chapter, we propose three different *Ga*-face configurations in order to perform a DC and RF fair comparison with the *N*-face configuration in terms of same gate-to-2DEG distance and capacitance.

We like to stress that the main aim of this work is not to offer optimized layouts for *Ga* and *N*-face HEMTs; here we rather try to compare devices built with two different technologies in order to identify the favorable aspects of each choice. The issue of how to compare devices based on different configurations of the same material is very similar to the one arising when attempting a comparison between devices built with different materials: finding a fair benchmark is an arduous task. We think that our approach succeeds in this task, but we are aware of the possibly controversial nature of some of our layout and testing choices. We first discuss the approach chosen for the comparison in Section 4.2. Then, we analyze the effects of the gate scaling on the DC and RF device performance in both technologies in Section 4.3. Conclusions are finally drawn in Section 4.4.

4.2 Comparison between N-face and Ga-face Technology

In order to investigate the advantages of the new *N*-face technology on the *Ga*-face one, we attempted a comparison of the DC and RF performance of both devices. The simulated *N*-face device is the one already described in Chapter 3, which we have experimental data for. As shown in Figure 4.1, the *Ga*-face structure used to perform

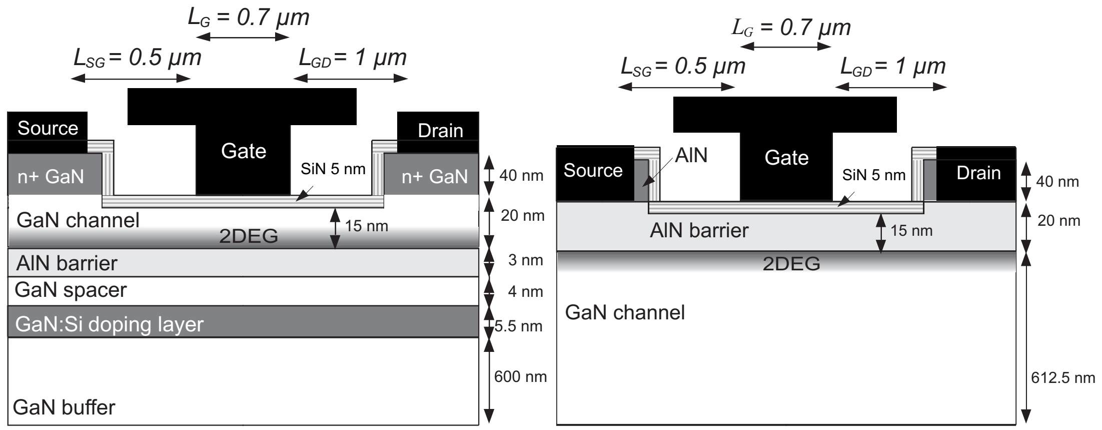

Figure 4.1: Comparison of the simulated *N*-face (left) and *Ga*-face (right) HEMT layouts

the comparison was directly derived from the *N*-face structure and kept as similar to it as possible: the *GaN* channel layer and the *AlN* barrier layer were swapped, the corresponding primitive vectors were rotated, and the *Si*-doped layer was removed. Furthermore, in order to obtain a fair comparison between the two different structures, we imposed the condition that the integral of the charge along the perpendicular direction to the channel is the same in both structures when the bias is $V_G = 0\ V$ and $V_D = 0\ V$. That is, we impose the constraint that both structures produce the same channel charge. Only in this way is the comparison fair.

As mentioned in the Introduction of this chapter, the presence of the *AlN* layer in the *N*-face configuration enhances the electron confinement in the channel with respect to the *Ga*-face one. Indeed, the large polarization-induced electric field in the *AlN* layer raises the *GaN* buffer conduction band edge with respect to the *GaN* channel, creating a potential barrier against carrier diffusion in the bulk layer. This property is highlighted in Figure 4.2, showing the carrier distribution (black points, at $V_{GS} = 0\ V$, $V_{DS} = 0\ V$) superimposed to the conduction band profile of the simulated devices. In Figure 4.3, a comparison is shown of the simulated I_D-V_D characteristics of the *N*-face (solid lines) and *Ga*-face devices (dashed lines). The I_D-V_G characteristic and the corresponding transconductance are illustrated in Figure 4.3. The *N*-face device exhibits a

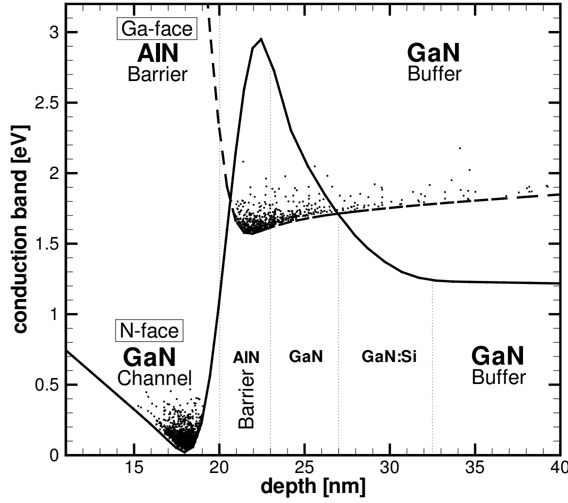

Figure 4.2: Conduction band profile, and carrier distribution (black points) corresponding to a bias of 0 V on the gate and 0 V on the drain of both N-face (solid line) and Ga-face (dashed line) devices. The higher carrier confinement in the N-face HEMT is evident.

higher transconductance peak of about 340 mS/mm for a gate voltage of $V_G = 4\ V$. As we can see, both the drain current and the transconductance provided by the N-face device are about 15% higher with respect to the one in the Ga-face configuration, which exhibits a peak transconductance of 300 mS/mm for the same gate voltage. However, the obtained peak transconductance values are overestimated, with respect to the experimental values (simulated 340 mS/mm versus experimental 240 mS/mm), due to the self-heating effects not taken into account in the present work [11], and the non-ideality related to real N-face material growth issues.

This first transconductance data can lead to the conclusion that the N-face device is much more efficient than the Ga-face device. However, the device performance depends on different aspects that are not correlated with the polarization type. In fact, the gate-to-channel distance and the dielectric constant of the upper layer play an important role in the device performance, and it is evident from Figure 4.2 that the N-face device has higher gate-to-channel capacitance.

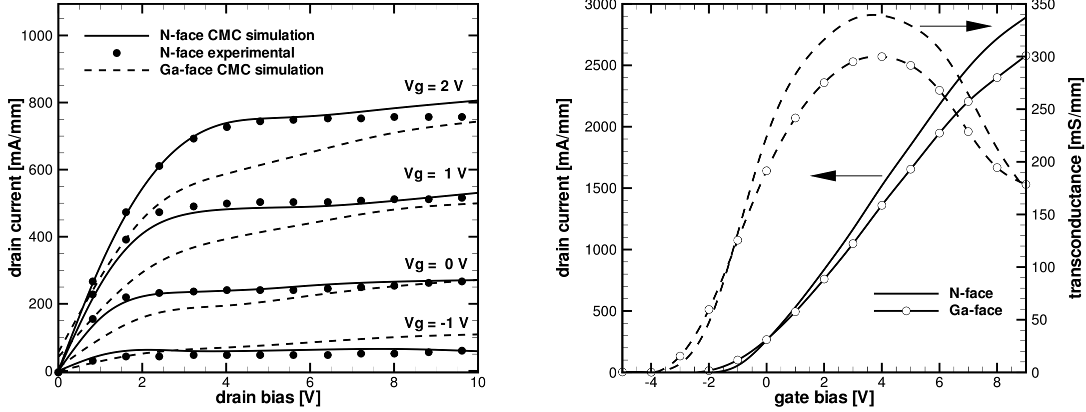

Figure 4.3: LEFT: Comparison of simulated (solid lines) and experimental (dots) [10] $I_D - V_D$ characteristics of the N-face HEMT. The CMC simulation $I_D - V_D$ of the Ga-face device is also shown (dashed lines). The gate voltage V_G is varied between 2 V and -1 V with a bias step of $\Delta V_G = 1$ V. RIGHT: Simulated I_D-V_G characteristic and corresponding transconductance for a drain bias of $V_D = 10$ V.

In order to take into account these features and obtain a fair comparison, we analyzed other two device configurations. In the first one, we added a second constraint to the gate-to-channel distance, designing the Ga-face to obtain the same vertical distance between the gate and the carrier density peak in the channel, similarly to the N-face device. Finally, we defined a second configuration that further reduces the gate-to-channel distance in the Ga-face device in order to have the same gate-to-channel capacitance as the N-face device. In this way, we mitigated the effect of the lower AlN dielectric constant. This last configuration allowed us to isolate the inherent back-barrier effect on the device performance.

The charge densities along the direction perpendicular to the channel for the three configurations considered are shown in Figure 4.4. The improved carrier confinement in the N-face configuration is also particularly evident in the figure. The contact resistance is kept constant in all configurations in order to reduce as much as possible the differences between the device structures studied.

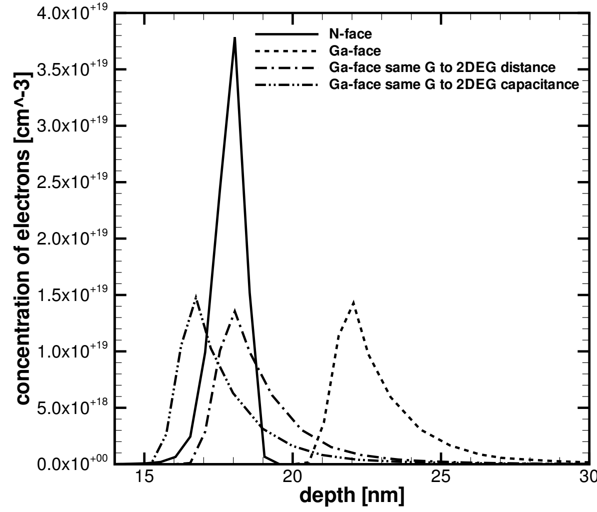

Figure 4.4: Charge density along the direction perpendicular to the channel for the three cases described in the text, for a bias of $V_G = V_D = 0\ V$. The integral of the charge is the same for all the four configurations shown.

The I_D-V_G characteristics, and the corresponding transconductance, of the *N*-face (solid lines) and *Ga*-face (dashed lines) devices, are illustrated in the upper panel of Figure 4.5. It can be seen that most of the current increase is due to the different channel-to-gate distance. The dielectric constant of the layer between the gate and the channel plays, however, an important role as shown by the curve related to the equal gate-capacitance configuration: setting the gate-to-channel distance in order to obtain the same gate capacitance yields a very similar transconductance peak values for both the *Ga*- and *N*-face configurations. In this case, only for high V_{GS} values, an improvement of the *N*-face device performance can be observed due to the better carrier confinement. Moreover, the *N*-face device exhibits a broader and flatter transconductance that will allow better linearity under large signal operations. Furthermore, when evaluating the drain current and the transconductance for a low drain voltage of $V_D = 5\ V$, we can see from the lower panel of Figure 4.5 that the *N*-face device performs better than any other *Ga*-face configuration.

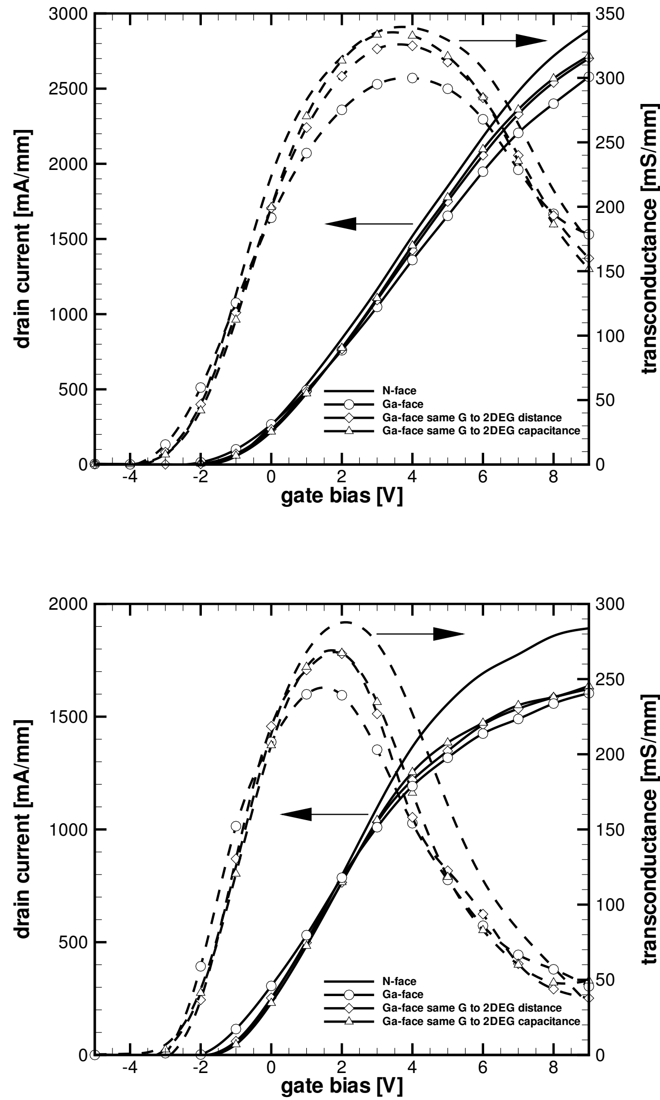

Figure 4.5: Simulated I_D-V_G characteristic and corresponding transconductance for $V_D = 10\ V$ (upper panel) and $V_D = 5\ V$ (lower panel). The *N*-face device and the three different *Ga*-face device configurations described in the text are compared.

It must be noted that the interface roughness, not included in the present work, may also play an important role when comparing the *N*-face and the *Ga*-face technology. Indeed, the negative gate bias of such depletion mode devices pushes the electrons away from the gate. This results in the fact that the electrons are pushed toward the back barrier (*i.e.* the *GaN/AlN* heterojunction) in *N*-face devices, while the electrons are pushed away from the barrier (*i.e.* the *AlN/GaN* heterojunction) in *Ga*-face devices. Thus, a larger scattering due to interface roughness is expected to occur in *N*-face device.

4.3 Back Barrier Effect on Device Scaling

Short gate lengths are usually coupled with the occurrence of short channel effects such as threshold-voltage shift, soft pinchoff, and high sub-threshold current, due to the poor electron confinement in the 2DEG channel, which also reduces the gate modulation efficiency [83] as further discussed in Chapter 5. In the *N*-face structure analyzed, the *AlN* forms an inherent back barrier for reducing short channel effects, forming a strong back-barrier confinement of the 2DEG at high electric fields, which facilitates the device scaling [73]. To understand the effects of the barrier layer on scaling, we analyzed the DC and the frequency response of the *N*-face and the *Ga*-face devices in the last two configurations, by varying the gate length.

Gate Length Scaling Effect on DC Performance

In order to investigate the effects of the gate length scaling on the DC performance, the $I_D - V_D$ characteristics for $L_G = 700\ nm$, $300\ nm$, and $150\ nm$ were calculated and are shown in Figure 4.6 for the two cases analyzed: *Ga*-face and *N*-face with the same gate-to-channel distance, and *Ga*-face and *N*-face with the same gate-to-channel-capacitance. As we can see, decreasing the gate length increases the drain current – and consequently the transconductance peak value – leading to an improved device performance. However, a threshold shift towards negative values, caused by the drain induced barrier lowering (DIBL) effect, is also detected. This trend is present in both

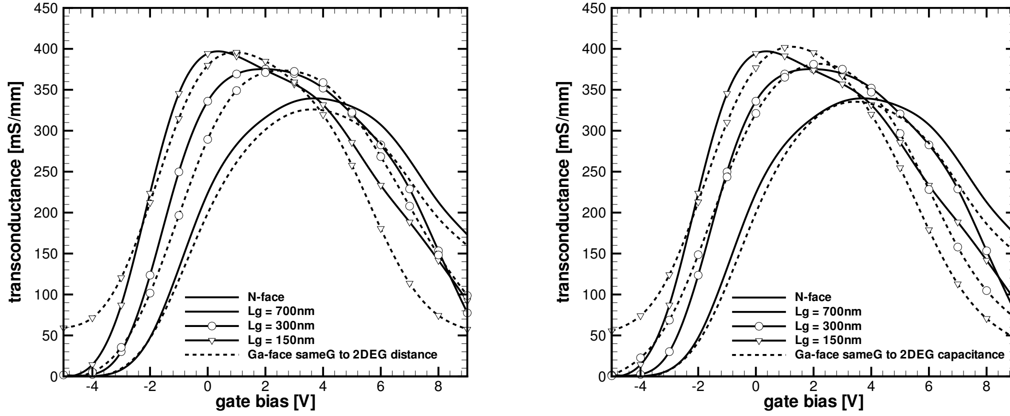

Figure 4.6: Simulated g_m-V_G curves for $V_D = 10\ V$ for different device gate length ($L_G = 700\ nm$ to $L_G = 150\ nm$). The *N*-face device (solid line) is compared with the *Ga*-face device (dashed line) with the same gate-to-2DEG distance (left) the same gate-to-2DEG capacitance (right).

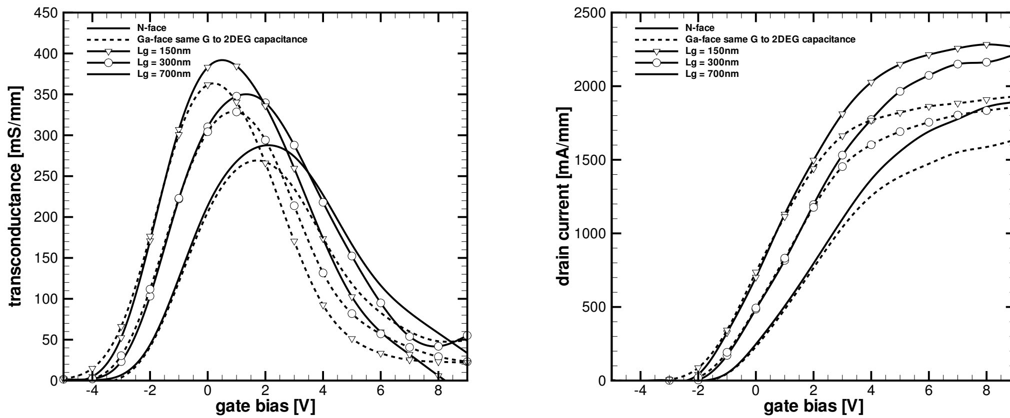

Figure 4.7: Simulated g_m-V_G (left) and I_D-V_G (right) curves for $V_D = 5\ V$ for different device gate length ($L_G = 700\ nm$ to $L_G = 150\ nm$). The *N*-face device (solid line) and the *Ga*-face device configuration with the same gate-to-2DEG capacitance (dashed line) are compared.

N- and *Ga*-face structures, even though this phenomenon is accompanied by an higher increase of the subthreshold current in the *Ga*-face devices. Moreover, from Figure 4.7 we can see that, for the low drain voltage of $V_D = 5\ V$, the *N*-face device performs better in terms of transconductance and maximum drain current, than the best *Ga*-face configuration (*i.e.* same gate-to-2DEG capacitance).

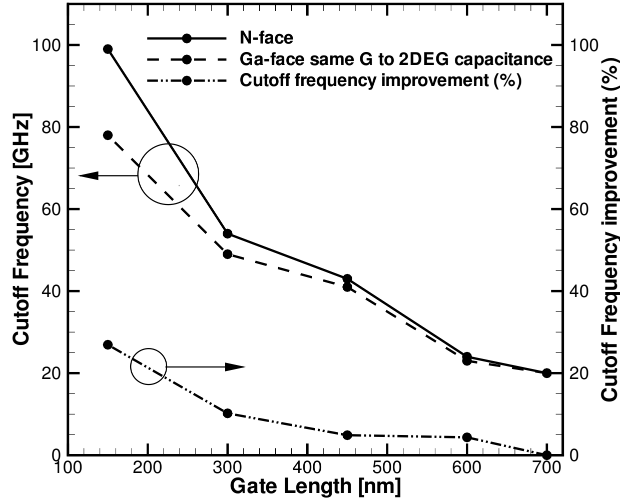

Figure 4.8: Cut-off frequencies of the *N*- (solid line) and *Ga*-face (dashed line) *GaN* HEMTs, with the same gate-to-2DEG capacitance, extracted by the CMC simulation performed for $V_D = 10\,V$ and a gate biased corresponding to the peak transconductance, with gate length from $L_G = 700\,nm$ to $L_G = 150\,nm$. The dashed dot line represents the cut-off frequency percentage improvement of the *N*-face respect to the *Ga*-face device, as a function of the gate length.

Gate Length Scaling Effect on RF Performance

To understand how the gate scaling influences the RF performance of *N*- and *Ga*-face devices, we analyzed the frequency response at different the gate lengths. In order to have a fair comparison, the *Ga*-face configuration with the same gate-to-channel capacitance was chosen for this analysis due to its similar DC characteristics with respect to the *N*-face device. In particular, the cut-off frequency is evaluated for a drain voltage of $V_D = 10\ V$ where the two devices exhibit almost the same peak transconductance. The cut-off frequency for different gate lengths is shown in Figure 4.8 for both the *N*- (solid line) and *Ga*-face (dashed line) configurations. The values were extracted through small-signal analysis as described in Chapter 2. As we can see, even though the DC performance of the two devices is similar at the chosen bias point, the larger carrier-confinement in the *N*-face configuration produces a larger cut-off frequency es-

pecially at shorter gate lengths. In particular, the f_T values range from about 20 GHz for $L_G = 700$ nm in both devices, to 99 GHz and 78 GHz for $L_G = 150$ nm for the *N*- and *Ga*-face, respectively. This behavior is highlighted by the dashed dot line, which represents the cut-off frequency percentage improvement of the *N*-face respect to the *Ga*-face devices, as a function of the gate length.

4.4 Conclusions

In this chapter, we reported a comparisons between *N*- and *Ga*-face structures were to evaluate the potential of the newer *N*-face technology. In our analysis, structures of the same gate-to-channel distance as well as same gate-to-channel capacitance were investigated. Simulation results indicated that the *N*-face device allows larger cut-off frequencies with respect to the *Ga*-face, especially at shorter gate lengths, primarily because the improved carrier confinement reduces short channel effects. In general, the most important conclusion of this study is that, when *N*-face and *Ga*-face devices are properly compared, the *N*-face devices are advantageous due to the improved carrier confinement obtained by this structure.

Chapter 5

Highly-Scaled HEMT Optimization for Improved DC and High-frequency Performance

5.1 Introduction

Short-channel effects set a limit to the performance achievable through aggressive device scaling, which aims to reduce the carrier transit time in the gate control region by using deep sub-micron gate lengths. At such gate lengths, indeed, short-channel effects heavily affect both the DC and the RF performance of the device [55, 25, 31].

Generally speaking, short channel effects in DC regime are related to threshold-voltage shift, soft pinchoff, and high sub-threshold current due to the poor electron confinement in the 2DEG channel, which also reduces the gate modulation efficiency [83]. On the other hand, when the gate length is heavily scaled, the RF performance is affected by the fact that the difference between the extension of the effective gate length and metallurgical gate length cannot be neglected anymore. This results in a cut-off frequency which is a sub-linear function of the gate length. A significant mitigation of these short-channel effects can be achieved by down-scaling the gate-to-channel distance, d, with the gate length, L_g, to maintain a high aspect ratio, L_g/d, [85, 55]. On the other hand, excessively decreasing d has serious limitations in terms of gate leakage and reduction of the density and the carrier mobility in the 2DEG channel. In particular, in N-face devices (see Chapter 3 and 4), the electrons are pushed toward the back barrier ($i.e.$ the *GaN*/*AlN* heterojunction) with reducing d, which can results in a larger scattering due to interface roughness. Thus, maintaining high L_g/d imposes stringent trade-offs for device engineers. Conventional wisdom predicts for *GaAs* devices a minimum L_g/d of 2.5 - 6 [86, 87], whereas a larger value of L_g/d on the order of 10 is found for *GaN* ones [88, 89, 90, 25, 31, 34]. However, Jessen *et al.* reported more recently an even higher value of 15 for *GaN* transistors [85].

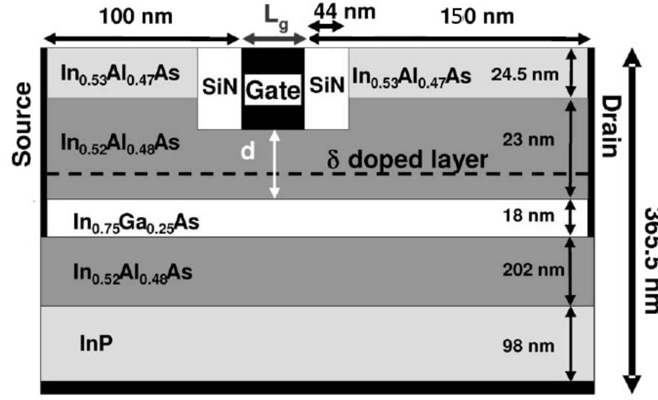

Figure 5.1: *InGaAs/InAlAs* HEMT with *InGaAs* quantum well channel simulated layout.

In this chapter, we study how the L_g and L_g/d scaling affect the RF and the DC performance of state-of-the-art HEMT devices, focusing on the two different *GaN* and *InGaAs* HEMT families and relating L_g/d to short-channel effects. The devices are characterized only in terms of intrinsic transistor gate area and access region resistance, as described in Chapter 2; furthermore, the effect of external parasitics is not taken into account here. The *GaN* HEMT considered in this study is the *N*-face device already described in Chapter 3, while the *InGaAs* device is a state-of-the-art *InGaAs/InAlAs* HEMT with an *InGaAs* quantum-well channel and terahertz capabilityance [30]. The *InGaAs* device layout in shown in Fig. 5.1 and is comprehensively described in [55] and [91], where it was showed that f_T can be significantly enhanced by scaling the gate length and gate-to-channel distance appropriately, and a theoretical upper bound for f_T exceeding 3 THz was established. In the next sections we report an optimization in terms of lateral and vertical scaling in order to maximize its frequency performance.

5.2 Effects of Aspect Ratio Scaling on DC performance

The impact of the aspect ratio on the DC performance was investigated by down scaling L_g and keeping constant d at 15 nm for both the *GaN* and the *InGaAs* devices. Moreover, the length of the access regions (*i.e.* source-to-gate and gate-to-drain spacing) were kept constant in order to preserve the same access resistance for each device.

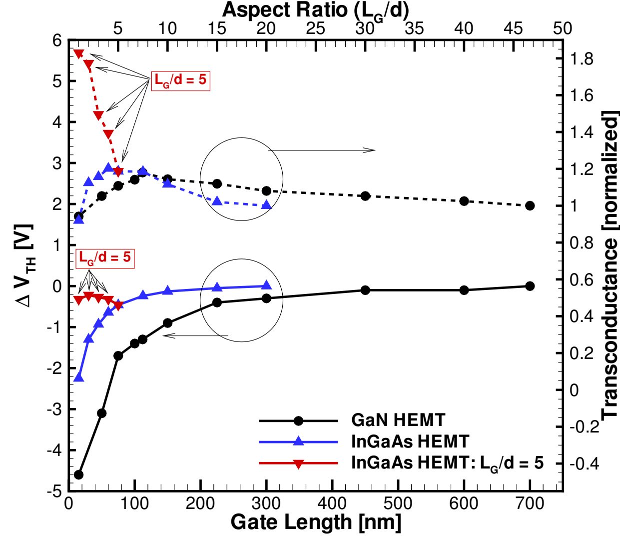

Figure 5.2: Comparison of DC short-channel effects for the *GaN* ($V_D = 10\ V$) (circles) and *InGaAs* ($V_D = 2\ V$) (triangles) devices. For each device material family, the device with the largest L_g/d was used as reference for ΔV_{th} and as normalization for g_m.

For each device family, the simulated device with the longest gate length is used as reference to understand how the scaling changes the DC characteristics of the devices with shorter L_g. This choice is due to the fact that short-channel effects are negligible in devices with longer gate lengths due to their high aspect ratios (> 20).

We first discuss the threshold voltage, V_{th}, behavior in *GaN* devices. The threshold voltage shift, ΔV_{th}, is shown in Figure 5.2 where devices with a high L_g/d values (down to a value of 15) show small variation of V_{th}. The decreasing L_g/d due to the L_g down-scaling leads to a shift of V_{th}, with respect to the V_{th} of the "reference" device with the longest L_g. A L_g/d value of 15 allows a ΔV_{th} less than 10% of the largest ΔV_{th} obtained in the *GaN* device with the shortest L_g ($L_g/d = 1$). At the same time, a ratio above 10 produces a ΔV_{th} value within 20% of the same largest ΔV_{th}. The observed ΔV_{th} is related to a reduced shielding of the gate from the drain field due to the shorter L_g and lower L_g/d as reported by Awano *et al.* [86], which results in a considerable influence of the drain field on the potential distribution on the source side of the channel.

Using an approach similar to the one used for the *GaN* devices, we evaluated the *InGaAs* devices. From Figure 5.2, we can see that *InGaAs* devices with a value of L_g/d down to 7.5 show significant ΔV_{th} that is about 10% of the largest ΔV_{th} obtained in the $L_g/d = 1$ *InGaAs* device. A ΔV_{th} value of about 15% of the largest ΔV_{th} can be observed for a ratio down to 5.

For comparison, a scaling of L_g with constant ratio L_g/d was also simulated for *InGaAs*. We maintained $L_g/d = 5$ for the shorter devices ($L_g < 75$ *nm*), allowing the gate-to-channel distance to scale down with the gate length. This was done by increasing the gate recess depth (*i.e* decreasing *d*). The results show how the negative ΔV_{th} is greatly reduced, and that in some cases it decreases with decreasing L_g (and related decreasing of *d*). This is due to both the reduction of short-channel effects (*i.e* reduced V_{th} negative shift), as well as to the depletion of the 2DEG due to the increased gate recess depth (*i.e.* gate closer to the 2DEG resulting in a V_{th} positive shift). Indeed, these two effects are combined here, leading to a complete elimination of the V_{th} negative shift.

The impact of aspect ratio and L_g scaling on the transconductance, g_m, were investigated as well for the sake of comparison. The g_m values for the different gate lengths were normalized with respect to the g_m value of the device with the longest L_g for each device family. The g_m results are also shown in Figure 5.2, where we verify that g_m increases with decreasing L_g. However, g_m starts dropping to values even lower than the reference one as soon as the gate loses the ability of modulating the channel due to short-channel effects and low L_g/d. In *GaN* devices, the maximum g_m is achieved for $L_g = 112$ *nm*, which corresponds to $L_g/d = 7.5$. On the other hand, g_m in *InGaAs* devices achieves larger normalized values, because the g_m peak occurs at shorter $L_g = 60$ *nm*, corresponding to $L_g/d = 4$. Observing the scaling with $L_g/d = 5$ in the *InGaAs* case, we can see that g_m is heavily enhanced by the decreasing gate-to-channel distance without showing reduced carrier mobility or excessive depletion in

the 2DEG channel in this case. The reduced short-channel effects due to the constant $L_g/d = 5$ scaling only plays a minor role in this particular situation.

These DC results, both for ΔV_{th} and g_m, are compatible with what experimentally found in *InGaAs* HEMTs by Kim and del Alamo [28], and in *GaN* HEMTs by Jessen *et al.* [85]. Moreover, our simulations allow us to directly identify aspect ratio requirements for the two different device families in terms of short-channel effect minimization: the *GaN* devices are more demanding in terms of aspect ratio than *InGaAs* devices.

5.3 Effects of Aspect Ratio Scaling on RF performance

The same gate length scaling, with constant gate-to-channel distance, described in the previous section was applied for the RF investigation. The cut-off frequency was extracted with small-signal analysis, and validated via a technique based on the electron weighted average velocity in the gate region. Both the extraction methods are discussed in Chapter 2. The obtained f_T values are surely overestimated, with respect to experimental results, due to the fact that we are only simulating the intrinsic transistor gate area and the resistance of the access regions without taking into account external parasitics, which heavily affect RF performance.

The frequency performance of *GaN* devices for different L_g values is shown in Figure 5.3, where the sub-linear behavior of f_T versus $1/L_g$ is evident. Similarly to the DC case, we proceed to quantify the RF short-channel effects, and this is done by evaluating the deviation from the linear f_T versus $1/L_g$ behavior. A ratio of 15 allows a deviation less than 5%, while a ratio of 10 allows a deviation within 15%.

The *InGaAs* RF results are shown in Figure 5.4. As we can see, f_T scales linearly down to a L_g/d ratio of 7.5 for *InGaAs* devices, and the f_T deviation from the linear value is within 15% for a value of 4-5. Below this L_g/d value, the device experiences a highly sub-linear f_T scaling behavior due to heavy RF short-channel

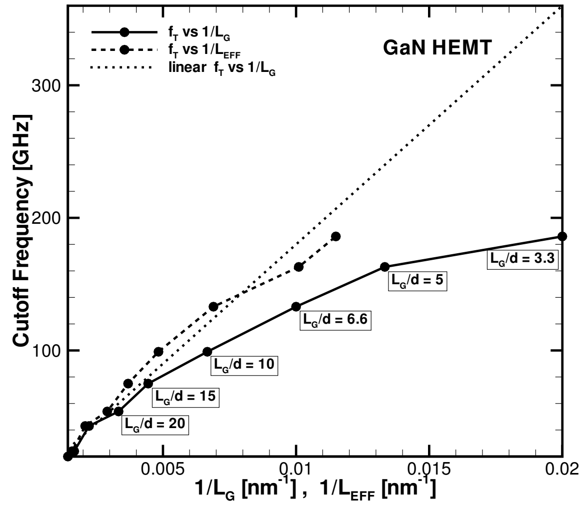

Figure 5.3: *GaN* HEMT cut-off frequency for $d = 15\ nm$ versus $1/L_g$ (solid line) and versus $1/L_{eff}$ (dashed line). V_g corresponds to the peak g_m, and $V_d = 10\ V$.

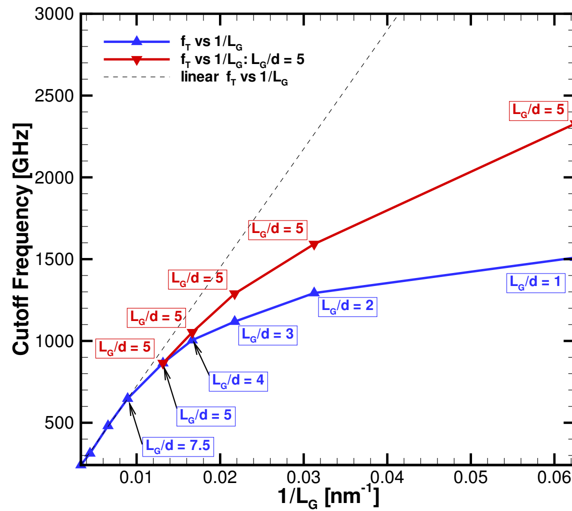

Figure 5.4: *InGaAs* HEMTs with fixed $d = 15\ nm$ (triangles) and fixed $L_g/d = 5$ (gradients) cut-off frequency versus $1/L_g$. V_g corresponds to the peak g_m, and $V_d = 2\ V$.

effects. A comparison between these last results and an *InGaAs* device scaling with constant L_g/d of 5 is also shown in Figure 5.4. In this configuration, we allowed d to be reduced with the reduction of L_g. As we can see, the device with constant aspect ratio performs much better in terms of f_T. Although the f_T sub-linear trend is still present, short-channel effects are noticeably reduced by preserving an acceptable high aspect ratio.

5.4 Role of Aspect Ratio and Effective Gate Length

The effective gate length, as defined in [55] and in Chapter 2, was further investigated and the plot of f_T versus $1/L_{eff}$ is shown in Figure 5.3 for *GaN*, and in Fig. 5.5 for *InGaAs*. The f_T value is found to scale linearly with respect to $1/L_{eff}$, regardless of the device family and whether the aspect ratio is kept constant or not. Therefore, this f_T versus $1/L_{eff}$ linear relation depends on the carrier velocity in the channel that is characteristic of the material, rather than on the particular adopted vertical geometrical configuration. In particular, the much larger *InGaAs* electron velocity with respect to the *GaN* one, results in a larger slope of this linear relation in *InGaAs* HEMTs with respect to *GaN* devices.

The sub-linear f_T scaling versus $1/L_g$ is due to the significant extension of the gate fringing fields in short-channel devices. Thus, the effective gate length becomes proportionally larger than the metallurgical gate length in highly-scaled devices, heavily affecting the carrier transit time. Observing the values of L_{eff} associated to devices with same L_g but different L_g/d in Fig. 5.5, we can see that L_{eff} is always smaller in the case with high ratio L_g/d. Moreover, observing the difference between L_{eff} and L_g in Figure 5.6, we notice that this difference remains constant as L_g is down-scaled in the case where a constant ratio L_g/d is preserved by allowing d to be down-scaled as well. Lastly, when comparing the values of L_{eff} associated to *InGaAs* and *GaN* devices with same L_g and fixed $d = 15$ *nm* (from Fig. 5.5 and Figure 5.3), we verify that the *InGaAs* HEMTs alway exhibit smaller L_{eff} than the *GaN* ones.

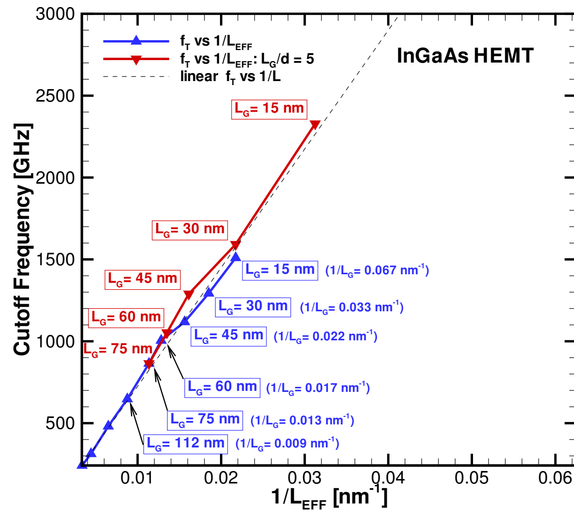

Figure 5.5: *InGaAs* HEMTs with fixed $d = 15\ nm$ (triangles) and fixed $L_g/d = 5$ (gradients) cut-off frequency versus $1/L_{eff}$. V_g corresponds to the peak g_m, and $V_d = 2\ V$.

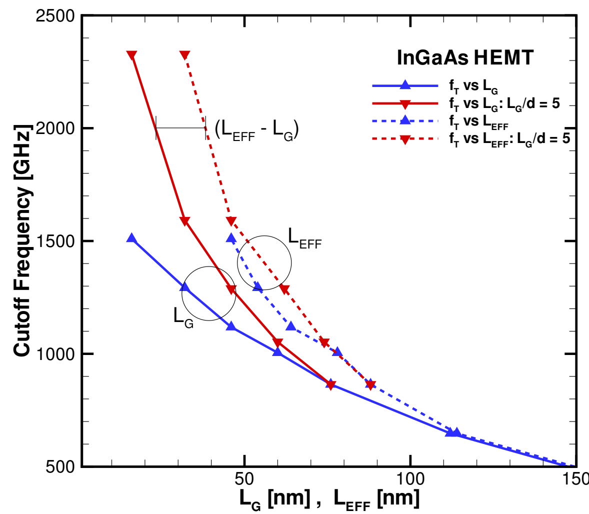

Figure 5.6: *InGaAs* HEMTs with fixed $d = 15\ nm$ (triangles) and fixed $L_g/d = 5$ (gradients) cut-off frequency versus L_g (solid line) and versus L_{eff} (dashed line). V_g corresponds to the peak g_m, and $V_d = 2\ V$.

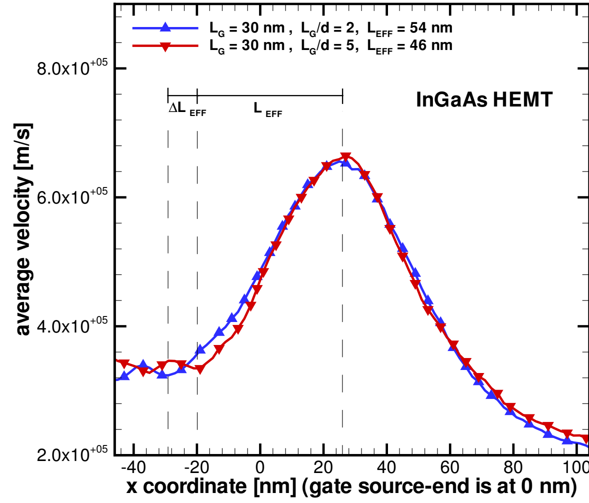

Figure 5.7: *InGaAs* HEMT with $d = 15\ nm$ (triangles), and *InGaAs* HEMT with a fixed aspect ratio of 5, and variable d, (gradients), average velocity in the channel region and L_{eff} for $L_g = 30\ nm$.

These phenomena can be explained by evaluating and comparing the electron velocity profiles in the channel. In Figure 5.7, we can see the velocity profile for two *InGaAs* devices, both with $L_g = 75\ nm$ but with different L_g/d. We first notice how the the velocity peak of $7.0 \times 10^5\ m/s$ is larger than the maximum bulk steady-state value, which from our *InGaAs* bulk simulations is found to be $2.4 \times 10^5\ m/s$ with an electric field of $3.1 \times 10^5\ V/m$. The bulk steady-state peak velocity is exceeded due to the quick acceleration in the gate region, and the related electron velocity time overshoot. Moreover, despite the velocity peak is the same for both the cases, L_{eff} is smaller for in device with larger L_g/d (*i.e.* smaller d). This is due to a reduction of the fringing field extension at the channel associated to higher L_g/d as reported by Jessen *et al.* [85] and reduced gate-to-channel distance.

The velocity profile of an *InGaAs* and *GaN* device with same $L_g = 75\ nm$ ($L_g/d = 5$) is compared in Figure 5.8. The *GaN* device exhibits velocity overshoot as well with a peak velocity of $3.5 \times 10^5\ m/s$, where the *GaN* bulk simulated max-

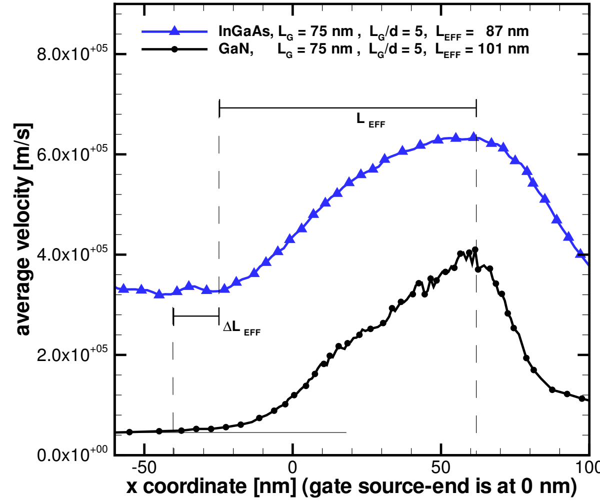

Figure 5.8: *InGaAs* HEMT with $d = 15$ nm (triangles), and *GaN* HEMT with $d = 15$ nm (circles) average velocity in the channel region and L_{eff} for $L_g = 75$ nm.

imum steady-state velocity is 2.5×10^5 m/s with an electric field of 2.7×10^7 V/m. So, despite the two materials have similar maximum steady-velocity, *InGaAs* devices can achieve much better frequency performance due to their higher transient overshoot velocity related to their much higher mobility. Furthermore, we notice that the *GaN* HEMT exhibits larger L_{eff} than the *InGaAs* at such low L_g/d value.

Moreover, the velocity in the access region is also much higher in the *InGaAs* device. This is due to the fact that the *InGaAs* superior electron mobility allows electrons to reach the velocity peak for lower electric fields than *GaN* (3.1×10^5 V/m versus 2.7×10^7 V/m). In the simulations, both devices had an access region electric field of the same order of magnitude (10^5 V/m). While this higher velocity does not directly translate in an improved f_T, which is only due to velocity in the gate control region, it allows a larger drain current. Anyway, if the electrons reach the acceleration point (*i.e.* the beginning of the gate control region) with a higher initial velocity, they will be able to achieve a higher overshoot transient peak velocity, with a positive impact on the transit time and f_T. In the *InGaAs* device, the access region velocity is slightly above

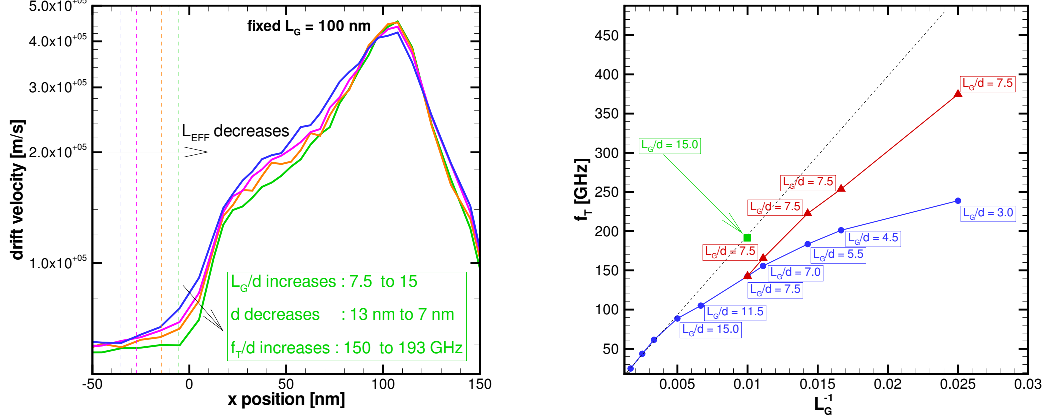

Figure 5.9: LEFT: Electron velocity profiles for same gate length but reducing gate-to-channel distance (*i.e.* increasing L_g/d). RIGHT: f_T versus L_G^{-1}. Downscaling d with L_g (*i.e.* constant $L_g/d = 7.5$) allows to improve f_T by reducing L_{eff}. The $L_g = 100\ nm$ device with decreased $d = 7\ nm$ and increased $L_g/d = 15$ is also shown in the picture as well as the theoretical linear scaling of f_T versus L_g^{-1} (dashed line).

the *InGaAs* bulk steady-state maximum values. In this particular device the source access region length, L_{sg}, is 104 *nm* and the carriers do not have time to slow-down once injected from the contact.

The relation between gate-to-channel distance and the effective gate length was also investigated for the $AlGaN/GaN$ HEMT with *InGaN* back-barrier described in Chapter 6. As we can see from Figure 5.9, a reduced spacing between the gate and the 2DEG channel reduces the extension of the fringing fields at the channel depth, delaying in such way the electron acceleration point and reducing the transit length (*i.e.* reduced L_{eff}). This results in a much smaller transit time and a consequent improved f_T. By increasing the aspect ratio, L_g/d, (*i.e.* decreasing d) from 7.5 to 15, the f_T of the device with $L_g = 100\ nm$ improved from 150 *GHz* to 193 *GHz*

Also for this device, we can assess the improved frequency performance when the gate-to-channel distance was down-scaled with the gate length as shown in the right side of Figure 5.9. As we can see, the f_T versus L_g^{-1} realtion deviates from the linear behavior for very short-channel devices. This is due to the increasing weight of

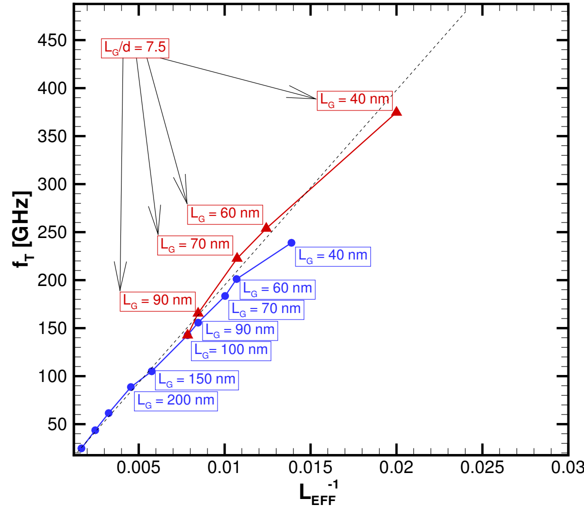

Figure 5.10: f_T versus L_{eff}^{-1}: this relation is linear regardless the adopted vertical geometry (*i.e.* value of d) and corresponds to the velocity characteristic of the material.

$(L_{eff} - L_g)$ with respect to L_g. However, by decreasing the gate-to-channel distance (*i.e* increasing L_g/d) we can reduce this deviation by containing L_{eff}. In particular, this deviation occurs as soon as L_g/d drops below 15 in GaN devices and, as previously discussed, we were able to improve f_T of the $L_g = 100\ nm$ device from 150 GHz to 193 Ghz, by using a reduced $d = 7\ nn$ corresponding to $L_g/d = 15$. This f_T value almost corresponds to the theoretical linear scaling value of 200 GHz achievable with this GaN device by down-scaling the device preserving a minimum aspect ratio for GaN-based devices of 15. On the other hand, it must be underlined that an excessively reduced gate-to-channel distance may result in fabrication issues, such as additional material defects due to a gate recess or reduced distance between surface traps and the 2DEG channel, and may not be always feasible in real device. Moreover, we can also see from Figure 5.10 that the relation between f_T and L_{eff}^{-1} is linear for this device as well in both cases with constant d and variable d. Thus, as discussed before, this linear relation is not dependent on the device geometry, but it is rather a property of the material and its slope is related to the material characteristic velocity.

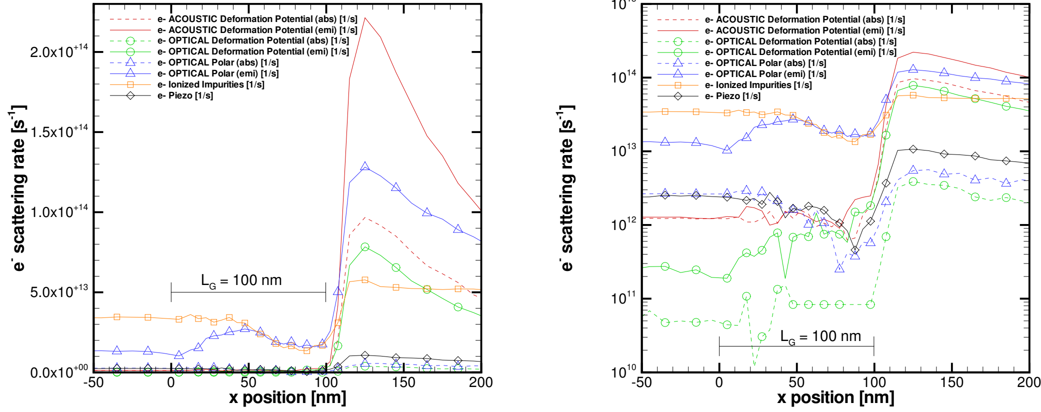

Figure 5.11: Separated scattering mechanism along the channel in linear (left) and logarithmic scale (right).

The contributions of the different scattering mechanisms limiting the electron velocity in the critical electron acceleration region (*i.e.* the source-side edge of the effective gate length) and along the channel are shown in Figure 5.11. This plot was obtained by simulating the device with the conventional Monte Carlo algorithm to obtain the information associated to the separate scattering mechanisms with the same approach described in Section 6.5. As we can see, the impurity scattering (due to the unintentional doping of $N_d = 1 \times 10^{15}\ cm^{-3}$ in GaN devices) is the dominant mechanism in the low electric field region located at the source-side of the drain. The optical phonon emission scattering also plays a role. On the other hand, as soon as the electrons transit past the drain-side edge of the gate, where the peak electric field is located (see Figure 6.5), the acoustic phonon deformation potential scattering suddenly becomes the dominant mechanism due to the high energy achieved by the carriers. We can also see a region, located between 70 and 100 nm, where the number of occurring scattering events is minimum due to the very low concentrations of electrons (as clear from Figure 6.3). In any case, past this region of minimum scattering, the increasing electron density and in particular the largely increasing electric field boost the scattering rate. Observing the low field region located at the beginning of L_{eff} in the same plot with a

logarithmic scale, we can see that the acoustic phonon deformation potential scattering is more than one order of magnitude smaller than the other low-field dominant scattering mechanisms. This is due to the fact that the acoustic phonon deformation potential is relevant only for high energy carriers.

5.5 Access Region Scaling for Improved Performance

A further improvement in ultra-short device performance can be achieved by downscaling the source-to-gate spacing, L_{sg}, wich results in a reduction of the access region parasitic source resistance (R_{si}). We investigated the impact of L_{sg} on g_m and f_T by changing L_{sg} between 104-674 *nm* in the $L_g = 76$ *nm InGaAs* device characterized in the previous section.

The L_{sg} scaling improves the DC performance in terms of transconductance according to the common relationship:

$$g_m = \frac{g_{m_0}}{1 + g_{m_0} R_{si}}, \tag{5.1}$$

and we can verify a large g_m improvement the simulated *InGaAs* HEMTs in Fig. 5.12. Moreover, improved high-frequency performance in terms of f_T for decreasing L_{sg} is also shown in Fig. 5.12. Observing Fig. 5.13, we verify how the increased L_{sg} does not affect L_{eff}, which is the same for all the L_{sg}. However, we notice a decreasing velocity in the gate region for increasing L_{sg}, which is the origin of the reducing f_T. However, lateral contacts located at the sides channel were used in the simulated devices. Whereas, in real devices, the contacts are located above the channel, and electrons are injected into the channel by tunneling through the *InAlAs* top cap layer. Therefore, the validity of these last results may be affected by this simplified simulated layout. Moreover, we did not perform a L_{sg} scaling in *GaN* devices in order to verify whether the same f_T improvement is possible. On the other hand, increased f_{max}, but with no effect on f_T, is reported for *GaN* devices with reduced access region lengths in [92].

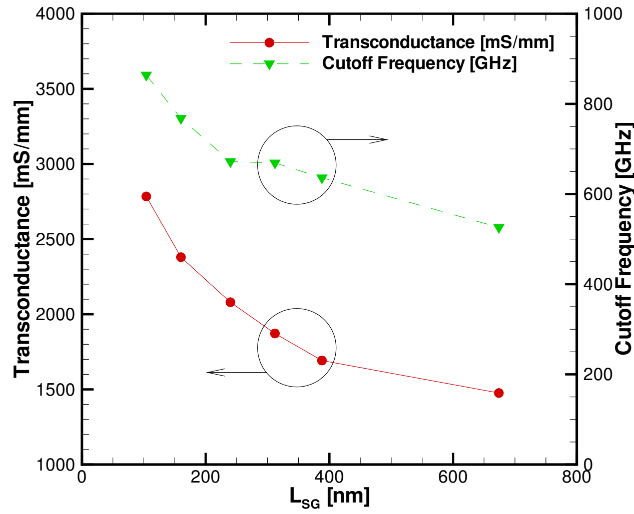

Figure 5.12: Transconductance (solid line) and f_T (dashed lines) for the *InGaAs* HEMT device L_{sg} scaling with $L_g = 76\ nm$.

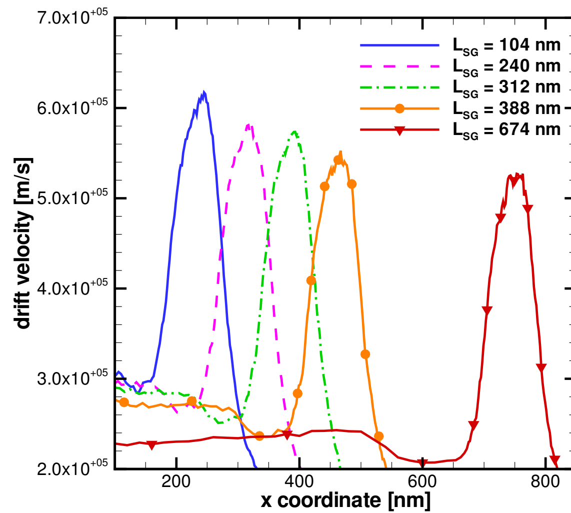

Figure 5.13: Average velocity in the channel region for the *InGaAs* HEMT device L_{sg} scaling with $L_g = 76\ nm$. The origin of the x-axis corresponds to the source.

5.6 Conclusions

We showed that when the gate length is down-scaled, L_{eff} becomes proportionally larger than the metallurgical gate length. This heavily affects low aspect ratio devices and is the critical factor limiting the RF performance of ultra-scaled devices. Moreover, maintaining a high aspect ratio prevents L_{eff} from becoming excessively larger than L_g, thus allowing improved RF performance by reducing the electron transit time. Moreover, the slope of the linear relation of f_T versus $1/L_{eff}$ was found to be associated to the device material family, and its characteristic carrier velocity in the channel, rather than to a particular vertical geometric configuration. Further DC and RF performance improvement for ultra-short devices can be achieved reducing the source-to-gate access region length, and the related parasitic access resistance. *GaN* HEMTs were found to be more demanding in terms of the minimum L_g/d than *InGaAs* devices. The larger L_{eff} of *GaN* HEMTs compared with *InGaAs* devices with same L_g suggests that the field distribution plays a major role in this picture. A shorter L_{eff} can be associated to weaker fringing electric fields, which are in turn related to the larger *InGaAs* HEMT relative dielectric constant in the layers between the gate and the 2DEG (12.5-14.5 in *InGaAs*-based devices versus 8.5-9.5 in nitride-based devices). Moreover, weaker fields reduce the drain field influence on V_{th}, and reduce ΔV_{th} in *InGaAs* HEMTs. This explains both the RF and DC reduced short-channel effects in *InGaAs* HEMTs.

We underline that *GaN* HEMTs with enhanced carrier confinement, such as *N*-face and *InGaN* back barrier devices, represent a ”best case” in terms of short-channel effects. Thus, conventional *GaN* HEMTs are expected to experience worse DC and RF short-channel effects once L_g/d falls below the reported minimum values. Moreover, we emphasize that our results, which do not take into account parasitics, represent an ideal upper limit in terms of performance for real devices. Anyway, they still provide meaningful guidelines for design aimed to device operating in the sub-millimeter band.

Chapter 6

Effects of the Passivation Dielectric Constant on the RF Performance of Millimiter-wave Power GaN HEMTs

6.1 Introduction

Large RF output power levels are commonly achieved in power *GaN* FETs through the increasing of the breakdown voltage by using field plates that redistribute the electric field [93, 94], and through the reduction of current collapse phenomena due to surface traps [95, 96] by using surface passivation. A field plate consists in an extension of the gate contact (in case of gate-connected field plates), or of the source contact (in case of source-connected field plates) [2], over the gate-drain region. Field plates reduce the voltage drop at the drain-edge of the gate, where the electric field peak is located, by forcing a potential value immediately past the gate, close to the potential of the gate. This results in the spreading of the electric field profile and a consequent reduction of the field peak value. The DC to RF current collapse causes the RF power of GaN-based HEMTs to be much lower than the one expected from the DC characteristics. This is manifested by a significant frequency slump or collapse in the RF drain current that severely reduces the output power and the power-added efficiency. The cause of this phenomenon is widely reported in literature [95, 96] as due to the presence of surface states between the gate and drain that deplete the channel with time constants long enough to disrupt the RF modulation of the channel charge. In particular, during large-signal operations, many electron are trapped in these surface states when the device output swing is at the extreme of the load-line, which corresponds to pinched-off channel and high drain voltage (*i.e.* high electric field at the gate drain-edge). In these conditions, trapped electrons are not able to contribute to the RF drain current (*i.e.* long de-trapping time constant) once the gate modulation opens the channel bringing the device to the other extreme of the load-line, which corresponds to open channel and low drain voltage. A smaller peak electric field due to the use of a field plate reduces

the device degradation due to the formation of additional traps by reducing the number of trapped hot-carriers. The current collapse phenomenon, experimentally characterized through pulsed measurements, is commonly modeled with the concept of virtual gate [96] that defines an additional depletion region, due to the negative charged surface traps, adjacent to the gate where the traps are usually located. The number of traps filled with electrons (*i.e.* the virtual gate potential) depends on the bias applied to the real gate. On the other hand, the slow response of the electrons trapped in the gate does not allow the virtual gate to respond fast enough to a high-frequency AC signal applied to the actual gate. The virtual gate concept must not be confused with the concept of effective gate length, discussed in Chapter 2 and further investigated in this chapter, that conversely defines the gate actual control region where a high-frequency AC gate signal is able to modulate the electrons in the channel.

SiN passivation is an established technology to effectively suppress surface states [97] and to provide mechanical protection for the device. It also provides a thick dielectric layer that enhances the effects of field plates. Other suitable dielectrics reported in literature as passivation materials are *BCB* [98], Al_2O_3 [99], Sc_2O_3 [100], and HfO_2 [101]. Moreover, due to the suppression via passivation of the negatively charged surface traps, the DC and RF performance also benefits from the increased charge density in the 2DEG, which in turn results in increased transconductance and reduced resistance in the access regions. On the other hand, the major negative effect of these dielectric layers with large relative dielectric constant is a significant increase of the gate fringing capacitances [85, 26, 25, 34] and parasitic capacitances. In particular, the drain-to-gate feedback capacitance, C_{gd}, heavily affects the RF performance of the device in terms of the maximum frequency of oscillation, f_{max}, due to the Miller effect that multiplies the C_{gd} impact on the input capacitance by a factor corresponding to the voltage gain [83]. Palacios *et al.* [12] reported improved small-signal high-frequency performance after removing the *SiN* passivation layer from the top of an $AlGaN/GaN$

device with a 25 *nm* barrier. More recently, by selectively removing the *SiN* dielectric from the immediate gate vicinity, Sun *et al.* [102] reported superior frequency performance with a very thin barrier (11 *nm*) $InAlN/GaN$ device. Moreover, the passivation dielectric constant and the trade-off between high-frequency versus high-power device design was described by Pei *et al.* [103], while the impact of *SiN* passivation on the small-signal parameters is reported in [104].

In this chapter, we further investigate these results by means of Monte Carlo simulations and we focus in particular on the role of the passivation layer dielectric constant with respect to the gate fringing fields, the parasitic capacitances, and the nano-scale carrier dynamics. We provide an understanding of the role of the *SiN* layer passivation independently from the surface trap suppression, and we show that there is in fact a trade off between the beneficial effects of a surface passivation layer and a degrading effect due to the change in electrostatics arising from the increased dielectric constant. We first describe the simulation method and the simulated device structure in Section 6.2. The effective gate length in relation to fringing capacitances is then investigated in Section 6.3. The carrier dynamics along the channel is further discussed in Section 6.4. The profiles of electric field and scattering for high-voltage drain biasing are shown in Section 6.5, while the related feedback capacitance and output impedance, obtained through small and large-signal analysis, are discussed in Section 6.6. Conclusions are finally drawn in Section 6.7.

6.2 Simulation Approach

The simulated device is a 100 *nm* gate length high-power *GaN* HEMT reported by Palacios *et al.* [12], with an *InGaN* layer grown between the *GaN* channel and the buffer layer, which acts as back-barrier. The simulated layout is shown in Figure 6.1. From bottom to top, a *GaN* buffer layer is followed by a 1 *nm* $In_{0.1}Ga_{0.9}N$ back-barrier, a 11 *nm* *GaN* channel, and a 25 *nm* $Al_{0.32}Ga_{0.68}N$ barrier. The gate is recessed in order to have a gate-to-channel distance of 13 *nm*. The source-gate and gate-drain access

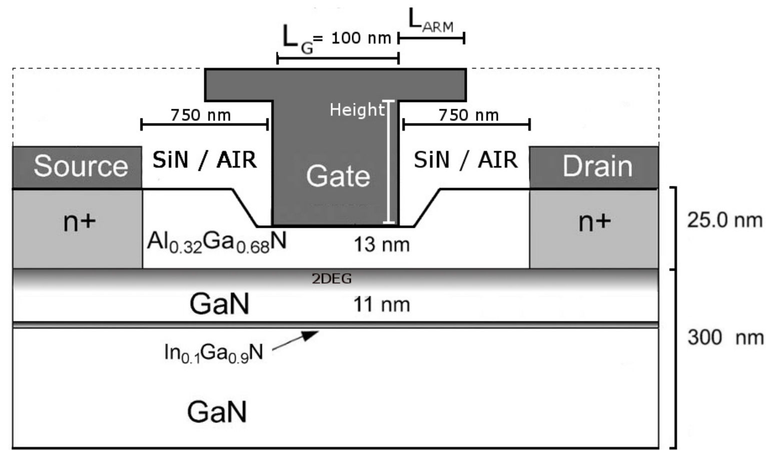

Figure 6.1: Generic schematic cross-section of the *GaN* HEMT simulation domain.

regions are both 0.75 μm allowing high-power operations. Further details about the simulator setup and the modeling approach can be found in [105]. In particular, the agreement between simulated and experimental DC/RF results for this specific device were previously reported by our group in [11] and is shown in Figure 6.2.

The main point of the present work is to show how the dielectric constant of the passivation layer impacts the electrostatics and the carrier dynamics, which strongly affects the frequency performance of the device. As such, we do not consider the reduction of surface traps between different passivation processes, as the specific purpose of the present work is isolating the role of the dielectric constant by comparing similar devices (*i.e.* same 2DEG density and same access region resistance). Furthermore, the comparison between a high-dielectric and low-dielectric constant can also be thought of as a simplified comparison between passivated and unpassivated (*i.e.* $\varepsilon_r = 1$) devices only in terms of the dielectric constant, where the role of the different surface trap densities is neglected. Within this scope, we investigate the difference between high-dielectric and low-dielectric constant layers, in terms of carrier dynamics (*i.e* carrier distribution and carrier velocity) and RF performance, by changing only the relative dielectric constant, ε_r, of the whole area around the gate outside the semiconductor as shown in Figure 6.1, and by conversely preserving the same polarization charge at

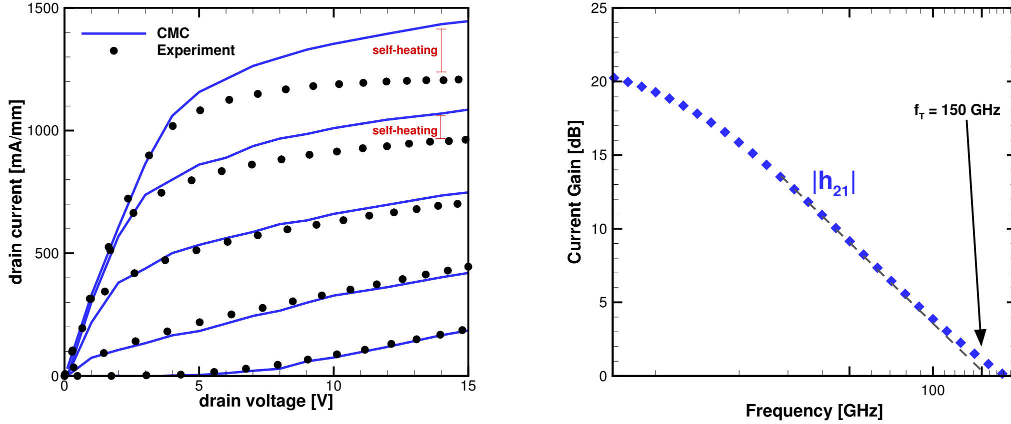

Figure 6.2: Agreement between simulated [11] and experimental [12] DC (left) and RF (right) characterization. The experimental f_T is 153 GHz.

the different interfaces within the device and at the surface. In particular, we focus on SiN passivation ($\varepsilon_r = 7.5$), but our results can be extended to any passivation material, where the electrostatic effects discussed here will be enhanced/reduced by a different ε_r. We also chose as reference for the low-dielectric constant case a value of $\varepsilon_r = 1$ ($i.e.$ air), which represents an absolute lower-bound. We are aware that the one offered here is not the complete picture of the many factors involved in the actual passivation process, but our contribution aims to underline and isolate the impact of a basic material property such as the dielectric constant in a complex parameter space in a manner not feasible experimentally.

6.3 Effective Gate Length and Cut-off frequency

The cut-off frequency and the effective gate length, L_{eff}, were obtained via the same technique based on the electron weighted average velocity in the gate region described in Chapter 2. The effect on L_{eff} of the gate fringing capacitances located at the lower edges of the gate was investigated by comparing a device with a dielectric constant surrounding the gate corresponding to air with one corresponding to SiN. In particular, we focus on a T-gate geometry with $L_{arm} = 0$ because the impact of the upper T-gate structure on the fringing electric field distribution located at the lower edges of the gate

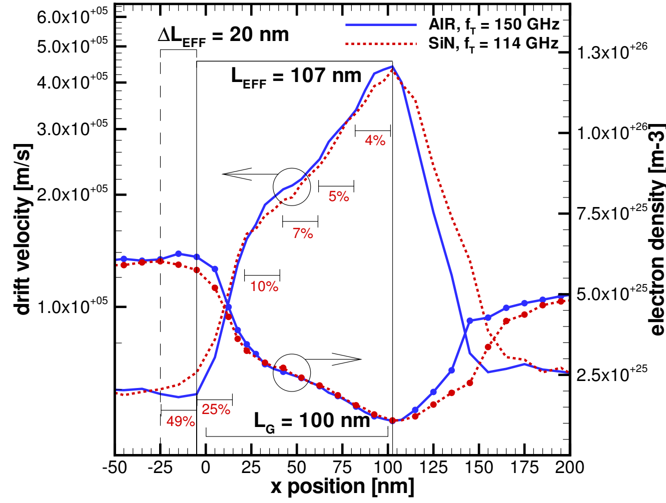

Figure 6.3: Comparison of the effective gate length extracted from the simulated electron velocity profile along the channel in the air (solid line) and *SiN* (dashed line) cases. The reported f_T is directly calculated from the transit time according to Eq.(2.16). The transit time percentage contribution due the different sections of L_{eff}, with respect to the total transit time, are reported in the *SiN* case. The electron density profile (circles) is also shown. The devices are biased for maximum f_T at $V_G = -2$ and V $V_D = 6$ V. The gate source-end is set as origin of the axis.

is found to be negligible. Moreover, we want to isolate the role of the gate from the role of the field plate, which can be thought of as secondary gate as discussed in the next sections. The devices were biased for their maximum f_T, which corresponds to V_G of -2 V and V_D of 6 V as experimentally reported by Palacios *et al.* [12]. As seen from Figure 6.3, the electrons in the *SiN* case start to accelerate about 20 nm sooner ($x = -25$ nm) than the air one ($x = -5$ nm). This is due to the fringing capacitances, which are enhanced by the larger dielectric constant of the *SiN* passivation. This extends the gate control region at the gate source-end and consequently extends L_{eff} with a detrimental effect on the total transit time. The velocity profile at the gate drain-end is also extended in the *SiN* case, but it does not affect our L_{eff} definition. Moreover, the position of the peak of the velocity is the same in both the devices at this bias point. We can also verify the estimate of the extended L_{eff}, in the source-side of the gate

area, from the electron concentration drop (*i.e.* depletion region) under the gate, which marks the beginning of the gate control region.

The difference in the velocity profile in the gate source-end region, in particular the starting point of the electron acceleration between the *SiN* and air cases, may at first appear negligible. However, from Eq.(2.16) and Figure 6.3, we observe that the slowest electrons, located in this region, have the highest impact on the transit time integral. Intuitively, the added transit time, due to a shift of the acceleration starting point, of the slowest electrons is much longer than the added transit time, due a shift of the peak velocity position, of the fastest electrons located near the velocity peak. In other words, the slow electrons at the gate source-end represent a "bottleneck" for the total transit time and the resulting in a substantial reduction of the frequency performance. These results were confirmed by estimating f_T also by small-signal AC analysis as described in Chapter 2.

SiN passivation is critical in modern *GaN* HEMTs for surface trap suppression and consequent current collapse reduction. We therefore simulated the same device progressively increasing the *SiN* layer thickness. We found that a *SiN* passivation layer thickness up to 5 *nm* yields a L_{eff} comparable with the same device with air. However, this is no longer true for a 15 *nm* thickness layer. These results indicate how the fringing capacitances are mainly affected by the dielectric constant of the material adjacent to the gate lower edges. This observation would suggest that a favorable trade-off between surface traps suppression and L_{eff} is possible by using thin *SiN* passivation layers. However, very thin *SiN* passivation layers are not effective in fully eliminating surface traps [106], suggesting that alternative passivation materials capable of effectively suppressing traps even when thin layers are employed, such as Al_2O_3 [106, 107, 99], are critical for future high-frequency high-power devices.

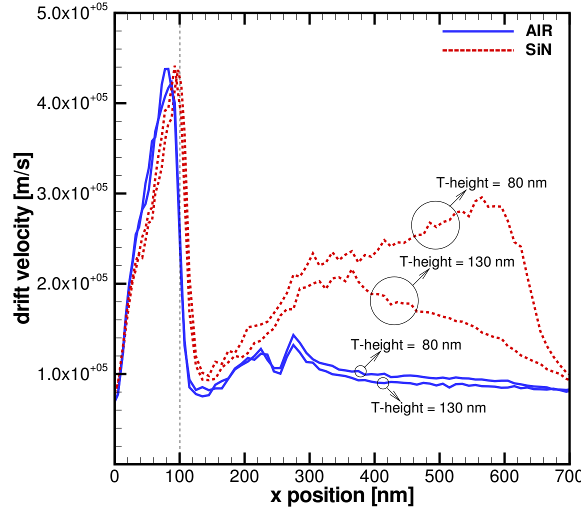

Figure 6.4: Velocity profile along the channel in the air (solid line) and *SiN* (dashed line) cases with $L_{arm} = 500$ *nm* and T-gate height of 80 and 130 *nm*. The devices are biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The vertical dashed lines represent the physical gate length drain-side end. The gate source-end is set as origin of the axis.

6.4 Carrier Dynamics along the Channel

The discussion of the carrier dynamics along the channel is critical to fully understand how important macroscopic device characteristics and phenomena such as frequency response and self-heating are affected by nanoscale carrier transport behavior. Thus, the devices simulated in the previous section were investigated in terms of their velocity-field relation, electron *k*-space distribution, and scattering rate profile along the channel. The devices were biased for their maximum f_{max} this time, which corresponds to a higher drain voltage V_D of 18 V as experimentally reported by Palacios *et al.* [12]. Moreover, this high V_D enhances the difference between devices with *SiN* and air as will be discussed later, and it is therefore helpful in our discussions.

We first discuss the role of the gate-connected field plate, which is usually thought of as a secondary gate in series with the actual gate. In the velocity-field curve of Figure 6.4, we can observe a second velocity peak located at a position along the

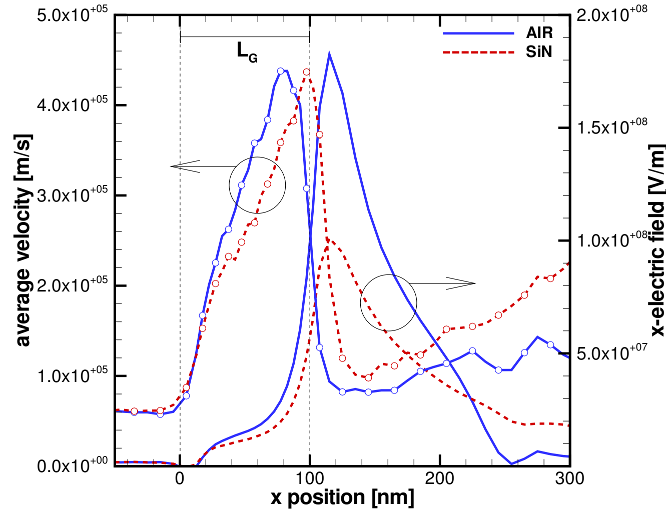

Figure 6.5: Electric field and velocity profile (dots) along the channel in the air (solid line) and *SiN* (dashed line) cases with $L_{arm} = 500$ *nm* and T-gate height of 80 *nm*. The devices are biased for maximum f_{max} at $V_G = -2$ *V* and $V_D = 18$ *V*. The vertical dashed lines represent the physical gate length (L_G) extension. The gate source-end is set as origin of the axis.

channel corresponding to the edge of the field plate. Moreover, a large dielectric constant of the passivation and a reduced T-gate height greatly enhance the control of the field plate on the channel. In particular, the second velocity peak corresponds to a second electric field peak along the channel, also located at the edge of the field plate. That field, in combination with a second depletion region under the field plate, effectively constitutes a second channel region under the control of the field plate. The drain voltage plays a crucial role in defining the carrier distribution in this second channel region. Specifically, the field, the velocity, and the depletion region are heavily dependent on the drain voltage, and they are much more reduced when a low potential of 6 *V* is applied (see Figure 6.14).

Accordingly to these observations, we chose the T-gate height of 80 *nm* and a field plate length, L_{arm}, of 500 *nm* as the reference case that best enhances differences between the *SiN* and air cases. With this configuration, we can see in Figure 6.5 how the

SiN large dielectric constant greatly enhances the field plate effectiveness. This because the larger dielectric constant allows the T-gate / field plate to more effectively spread the electric field distribution, and lowers the peak value as a consequence. A 45% reduction of the peak electric field can be observed in this case due to the enhanced control of the field plate, which in turn reduces the voltage drop at the edge of the gate. The dependence of the electric field on the T-gate dimensions is further discussed in the next section.

The velocity profile and the carrier dynamics can be further investigated from these results as well. We first notice that the velocity peak occurs 20 to 40 *nm* before the electric field peak in both cases. This occurs because the critical electric field, at which a significant fraction of electrons transfer to higher mass satellite valleys, is consistently lower than these electric field peak values. Thus, electrons achieve the peak velocity, and start to slow down due to negative resistance effect, well before the electric field reaches its peak. This also may explain why both devices exhibit the same peak velocity despite much different peak electric fields, being the peak velocity a material property, rather than a device property. The satellite valley population has to be only a few percent of the total population in order to have velocity saturation [108]. Moreover, the carriers in the gate region are accelerated by the rapidly increasing electric field up to the out-of-equilibrium peak velocity overshoot value, and they do not have time to slow down to the equilibrium lower maximum velocity value, which is 2.5×10^5 m/s in *GaN* bulk simulations. The same considerations hold for the results shown in the previous section with a lower V_D of 6 V. However, unlikely the previous section results, as shown in Figure 6.5, there is a 20 *nm* difference in the velocity peak position between the *SiN* and air cases with $V_D = 18\ V$. This is due to the fact that at such high V_D, the difference in terms of electric field profile are much more pronounced. So, the critical electric field in the device with *SiN* is reached later in the channel, as is clear from the plot, and the velocity peak is shifted as a consequence. However, this velocity

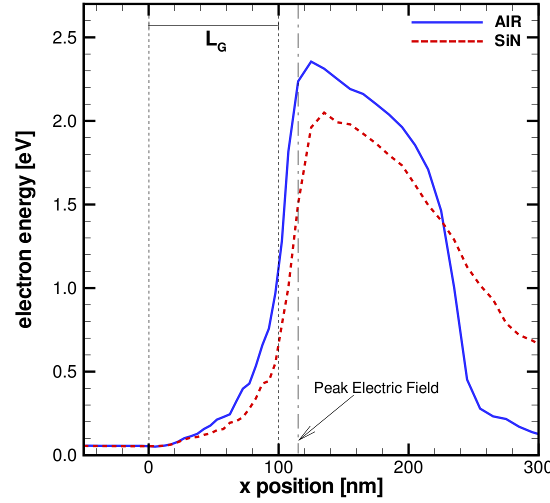

Figure 6.6: Electron energy profile along the channel in the air (solid line) and *SiN* (dashed line) cases with $L_{arm} = 500\ nm$ and T-gate height of 80 nm. The devices are biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The vertical dashed lines represent the physical gate length (L_G) extension, and the vertical dash-dotted line corresponds to the peak electric field position. The gate source-end is set as origin of the axis.

peak shift is not particularly critical in terms of frequency performance due to the high velocity of the electrons in this region, which is related to a short transit time compared to the total transit time as discussed in the previous section.

The electron energy, and the k-space electron distribution in the BZ, in relation to the position along the channel, have been investigated as well. Observing the electron energy profile in Figure 6.6, we notice that the energy does not decay as sharply as the electric field does at such large drain bias. Moreover, its value is still consistently large (about 2 eV) even at positions in the channel where the velocity has almost completely dropped and the electric field has decreased to relatively low values (*i.e.* 180 nm). This is because the electrons remain in the high-energy high-mass satellite valleys because of the large difference in density of states with respect to the Γ-valley, which lowers the probability to scatter back to lower energies. In order to verify this statement, we defined several 1-mesh-cell-length regions along the channel where we individually

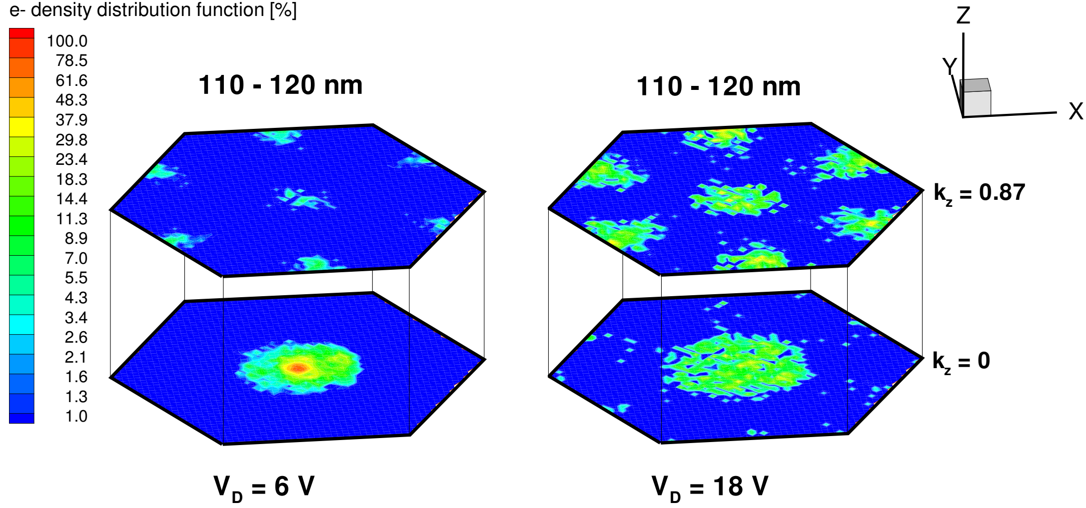

Figure 6.7: Average carrier distribution density in the momentum space at a position along the channel corresponding to the electric field peak in a device with air. The electron density is normalized to the maximum value within each single averaging region along the channel. The 3D Brillouin zone (normalized extension in the k_z-direction -1 to $+1$) was sliced with a plane corresponding to the minimum of the Γ-valley ($k_z = 0$), and with another plane corresponding to the minimum of the satellite valleys ($k_z = 0.87$). The devices are biased for maximum f_T at $V_G = -2\ V$ and $V_D = 6\ V$ (left), and maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$ (right). The gate source-end corresponds to the position $x = 0$.

computed the average k-space electron distribution. In such a way, we were able to track the evolution of the momentum distribution along the channel with a 5 to 10 nm resolution (depending on the local mesh size). The region extension along the vertical direction included the entire thickness of the channel layer in order to take into account all the electrons involved in the transport.

As we can see from Figure 6.7 in relation to a device with air, the large effective mass satellite valleys are significantly populated in correspondence to the gate drain-edge where the electric field peak is located. Moreover, the larger electric field at larger drain bias ($V_D = 6\ V$ and $E_{x_{peak}} = 0.8 \times 10^8\ V/m$ versus $V_D = 18\ V$ and $E_{x_{peak}} = 1.8 \times 10^8\ V/m$) results in a larger satellite valley population. The mapping along the channel at $V_D = 18\ V$ can be seen in Figure 6.8, where the k-space distribution at

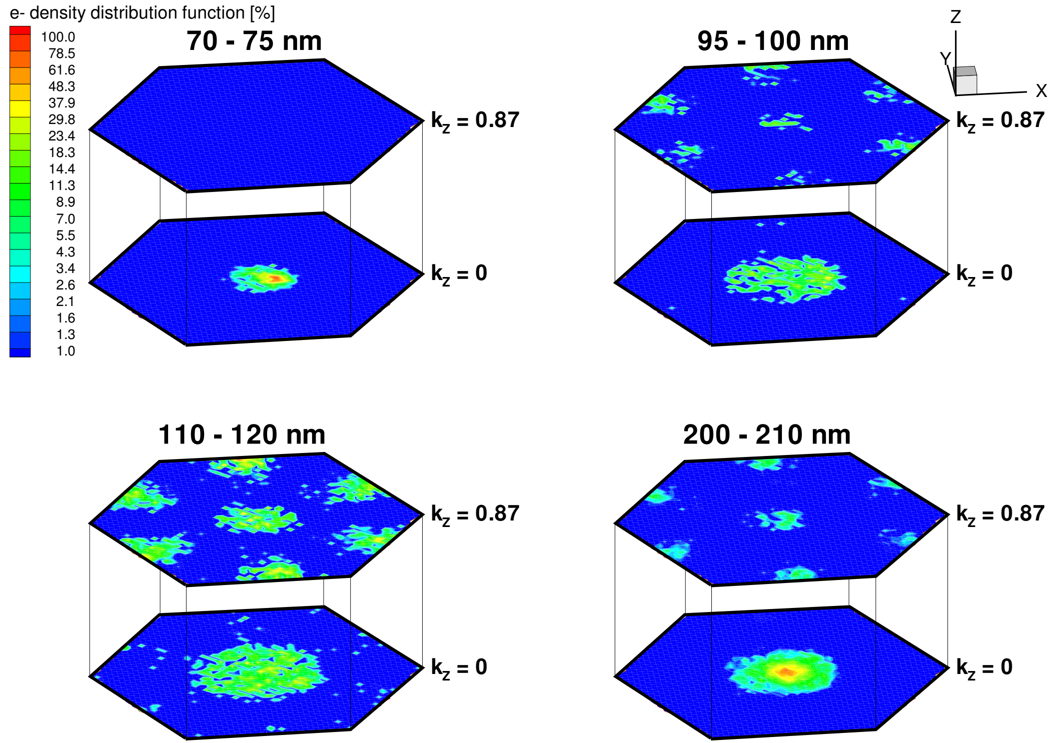

Figure 6.8: Average carrier distribution density in the momentum space corresponding to different positions along the channel in a device with air. The electron density is normalized to the maximum value within each single averaging region along the channel. The 3D Brillouin zone (normalized extension in the k_z-direction -1 to $+1$) was sliced with a plane corresponding to the minimum of the Γ-valley ($k_z = 0$), and with another plane corresponding to the minimum of the satellite valleys ($k_z = 0.87$). The device is biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The gate source-end corresponds to the position $x = 0$.

various relevant positions along the channel is shown. We clearly see that a significant number of electrons, initially completely distributed within the Γ-valley, are already transferred in the satellite valleys at a x-position corresponding to the velocity peak ($x = 90\ nm$). Then the satellite valley population achieves its maximum corresponding to the position of the peak electric field ($x = 120\ nm$). Finally, we can verify how the population of the satellite valley is still significant even 100 nm past the electric field peak ($x = 200\ nm$). In this position, the electric field intensity is about a quarter of the peak value. The satellite valley population eventually starts to drop around $x = 300\ nm$ and it is completely gone around $x = 400\ nm$. For the $V_D = 6\ V$ bias, the drop of the

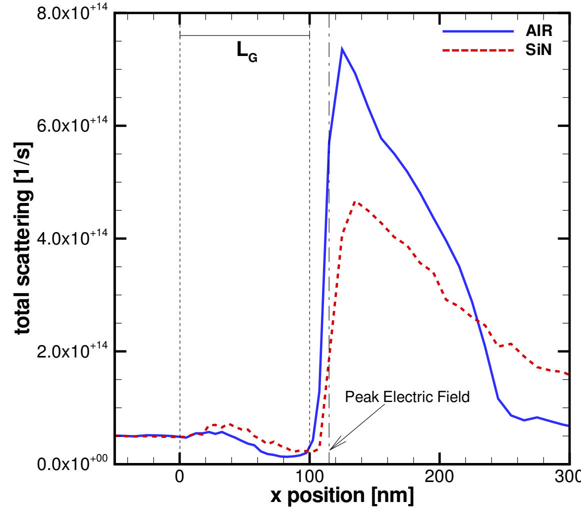

Figure 6.9: Electron scattering profile along the channel in the air (solid line) and *SiN* (dashed line) cases with $L_{arm} = 500\ nm$ and T-gate height of 80 nm. The devices are biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The vertical dashed lines represent the physical gate length (L_g) extension, and the vertical dash-dotted line corresponds to the peak electric field position. The gate source-end is set as origin of the axis.

.

satellite valley population occurs much sooner, around $x = 170\ nm$, due to the lower electric field.

The electron scattering profile is shown in Figure 6.9, and we notice how the peak value occurs past the electric field peak. This is related to the finite relaxation time of electrons, which determines this delayed scattering peak. Moreover, the observed scattering peak reduction between the *SiN* and air cases is proportionally smaller than the difference in the peak electric field. The difference of peak position and peak value increases with increasing drain voltages, and it is important in relation to self-heating and thermal-management issues in high-power device modeling. The joule heating is usually obtained as the dot product of the electric field and current density [109]. However, self-heating effects are related to electron-lattice interaction (*i.e.* phonons), and this delayed scattering peak, where phonon scattering is a large fraction, would

result in a hot spot located past the peak-electric field. Thus, a proper thermal modeling has to follow an electron-phonon dynamics approach in order to take into account the microscopic phonon emission non-locality near a strongly peaked electric field [110, 111, 112, 113].

6.5 High-power Biasing

The combined effect of the passivation dielectric and the T-gate on electric fields, scattering, electron density, and feedback capacitance were investigated by comparing the *SiN* and air cases with variable T-gate dimensions. The devices were biased with a large drain voltage of $V_D = 18\ V$, which is usually required to keep the device in the saturation region, accounting in such way for the load-line drain voltage swing, under large signal operations in millimeter-wave power amplifiers. Moreover, such large drain voltage corresponds to the device bias for maximum f_{max} as previously noted.

We first investigated the effect of the arm extension, L_{arm}, on the electric field. L_{arm} was varied between 0 and 500 *nm*. The effect of L_{arm} is minimal in the air case. Moreover, the T-gate height of 80 -*nm* was again chosen as a reference case that best enhances the differences in the *SiN* case, as discussed in the previous section. The electric field simulated at the $V_D = 18\ V$ bias is shown in the left side of Figure 6.10. We first notice how the peak electric field is reduced in presence of the *SiN*, and this occurs even in absence of an actual field plate structure (i.e. $L_{arm} = 0\ nm$). This is due to the fact that the gate contact is still able to provide a field plate effect through the lateral electric field originated from its sides. Increasing L_{arm} reduces the peak electric field, but this reduction behaves sub-linearly with respect of L_{arm}. In any case, we notice how increasing L_{arm} extends the control of the field plate thought of as secondary gate, as discussed in the previous section.

In order to verify the electric field distribution at the extreme of the load-line (see Section 2.5), which occurs during the large signal RF swing (quiescent point of $V_G = -2\ V$ and $V_D = 18\ V$), we also characterized the configurations just discussed at

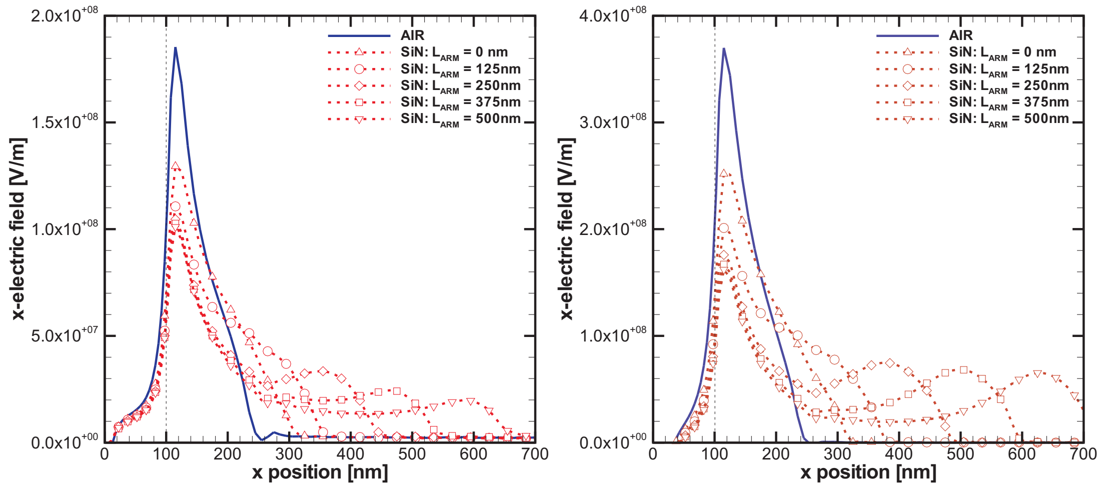

Figure 6.10: Electric field profile along the channel versus different T-gate arm lengths (L_{arm}) for the air (solid line) and SiN (dashed line) case with a T-gate height of 80 nm. The devices are biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$ (quiescent point) (left) and $V_G = -7\ V$ and $V_D = 28\ V$ (extreme of the load-line) (right). The vertical dashed lines represent the physical gate length (L_g) extension. The gate source-end is set as origin of the axis.

$V_G = -7\ V$ and $V_D = 28\ V$ (load-line extreme with low current and high voltage). As shown is in the right side of Figure 6.10, the maximum value achieved by the peak electric field is much larger in this second case ($1.8 \times 10^8\ V/m$ versus $3.8 \times 10^8\ V/m$), and there is a relatively more pronounced secondary electric field peak ($3.0 \times 10^7\ V/m$ versus $7.0 \times 10^7\ V/m$). In particular, by observing the area of the electric field profiles (*i.e* the voltage drop), we see that the difference between the voltage drop across the field plate (*i.e.* past 250 nm) and the voltage drop across the gate drain-end is proportionally larger the second case, corresponding to the extreme of the load-line. Therefore, the larger is the drain voltage, the larger percentage of the total voltage between gate and drain drops across the field plate.

The scattering rate profiles corresponding to the same set of simulations are plotted in Figure 6.11. The peak value of the scattering is not heavily affected by L_{arm} once this length is above 125 nm. However, L_{arm} does make a difference in terms of the drain-side tail of the scattering profile located in the second channel region controlled

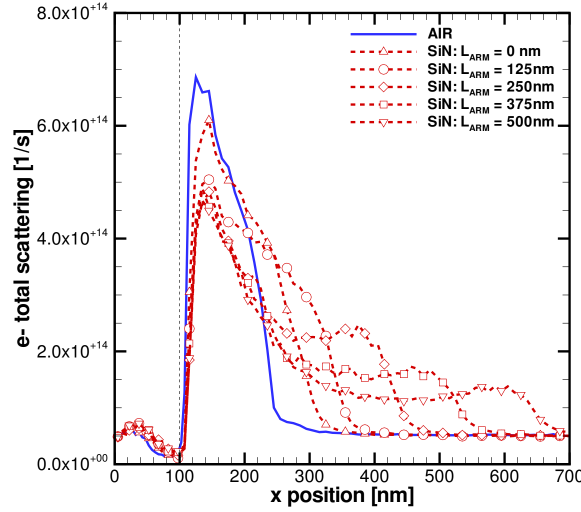

Figure 6.11: Electron scattering profile along the channel versus different T-gate arm lengths (L_{arm}) in the air (solid line) and SiN (dashed line) cases with a T-gate height of 80 nm. The devices are biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The vertical dashed lines represent the physical gate length drain-side end. The gate source-end is set as origin of the axis.

by the field plate. Keeping in mind the association between scattering and device self-heating discussed in the previous section, the scattering profile can be related to the extension and magnitude of the hot-spot that would occur in an experimental device due to self-heating. Despite a moderate effect on the peak value, further increasing L_{arm} would effectively spread the hot-spot, reducing the extension of a high temperature region (associate to high phonon emission), but conversely increasing the extension of the lower temperature region (the side tail of the scattering profile with lower scattering). With $L_{arm} = 125\ nm$ the scattering is still above a rate corresponding to the 60% of the peak value over a region extended 200 nm past the peak position. However, by increasing L_{arm}, the scattering profile drops below 60% in less than 100 nm.

The traditional EMC algorithm, described in Section 1.1, was employed to obtain a plot of the separate scattering mechanisms. In particular, the much faster CMC that does not allow the identification of individual scattering mechanisms since they

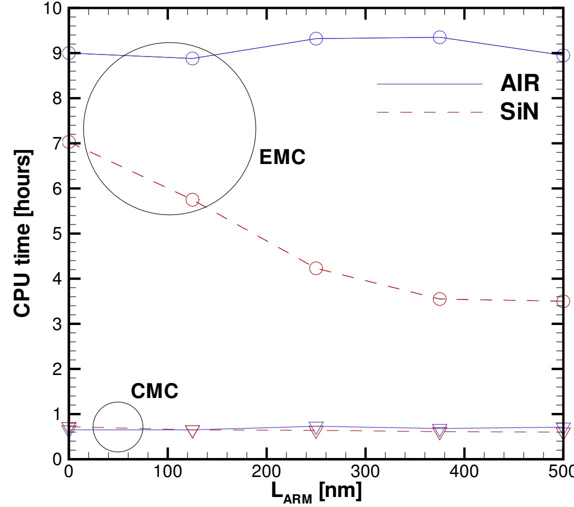

Figure 6.12: Comparison between computational time required by the CMC (triangles) and EMC (circles) algorithm to simulate additional 6 *ps* of the bias point corresponding to the maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The bias point steady state condition was formerly achieved through a 84 *ps* simulation by using the *CMC* algorithm. The simulations were run on a 64-bit Intel Xeon 3 *GHz*.

are all pre-tabulated [39], was used to simulated the device up to the steady-state corresponding to the desired drain voltage of 18 V. Simulating the device for a progressively increasing V_D from 0 to 18 V for a total simulated time of 84 *ps*, with 50000 carriers and a freeflight time step [38] set to 0.2 fs, required about 9 hours of actual computational time on a 64-bit Intel Xeon 3 *GHz* using 1.3 *Gb* of RAM to store the precomputed scattering look-up tables. Then,the position in the real and phase space of each individual carrier produced by the CMC was used as starting point of an additional EMC simulation of 6 *ps* . In this way, the computational burden was greatly reduced for the ECM. Quantitatively, this additional 6 *ps* EMC simulation took 4 to 9 hours depending on the device configuration, while the CMC required only about 40 minutes.

A comparison of the required computational time for the different configurations is shown in Figure 6.12. Observing the plot corresponding to the *SiN* case, we notice how the increasing scattering peak due to shorter L_{arm} almost doubles the EMC

computational time, while the CMC one is only slightly increased. Furthermore, the computational time for EMC is further increased by the larger scattering peak in the air case. In this last case there is no dependence on L_{arm}. This is due to the fact that the absence of the high dielectric layer decreases the effectiveness of the field plate effect as discussed before. Thus, it appears clear how the configuration with the largest scattering peak values, air with $L_{arm} = 0$, has a heavy impact on the EMC performance due to the large number of scattering events, which are not simulated as efficiently as the CMC does. In this case, the CMC is up to 14 times faster than the conventional EMC algorithm. The performance ratio can be even larger depending on the field distribution, and a ratio of 25 is reported for high-field bulk simulations [39].

The plot corresponding to the different scattering mechanisms with the device with air is shown in the left side of Figure 6.13, where the acoustic modes are associated to the three lowest eigenvalues of the dynamical matrix [40]. As it can be seen, the emission of acoustic deformation potential phonons is the dominant mechanism in the high-field gate-edge region, but then it quickly drops in the region where the electric field drops to low values. Comparing these results with the *SiN* case shown in the right side of Figure 6.13, we notice that the peak electric field reduction heavily affects primarily the acoustic deformation potential scattering, while the peak values of other relevant mechanisms are only partially reduced. This because the acoustic deformation potential is the major scattering mechanism only for high-energy carriers due to the density of states, and bulk simulations indicated that the associated scattering rate sharply increases at energies above 2 eV. But, at lower energy, other mechanisms such as polar optical emission and ionized impurity scattering are the dominant ones. In particular, the ionized impurity scattering shows a fairly constant profile both in the high-field region and in the low-field region past the gate edge (which corresponds to the second channel region controlled by the field plate as discussed before). This is due to the fact that the Coulomb interaction of electrons with the ionized impurities

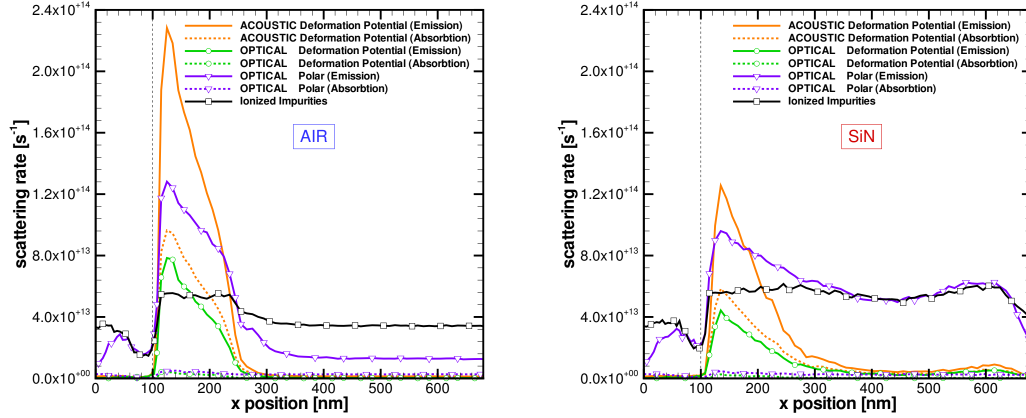

Figure 6.13: Air (left) and *SiN* (case) case EMC simulation plot of the different electron scattering mechanism profiles along the channel with $L_{arm} = 500$ *nm* and T-gate height of 80 *nm*. The acoustic modes are associated to the three lowest eigenvalues of the dynamical matrix. The devices is biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$. The vertical dashed lines represent the physical gate length drain-side end. The gate source-end is set as origin of the axis.

weakly affects high-speed high-energy carriers, but it is the dominant mechanism at very low energies (up to 0.15 eV). Furthermore, the extended drain-side tail of the electric field and energy, due to the enhanced spreading effect of the field plate in the *SiN* case, increases in the drain-side region the scattering mechanisms that are dominant for low/middle energy range. This explains why the polar optical emission, which is dominant in the middle-energy range (0.15 eV to 2.0 eV), dramatically increases over nearly the entire channel when the device has *SiN*. Over this same energy range, the ionized impurity scattering is still significant but not dominant.

Optical phonons are relatively slow (group velocity of about 1000 m/s), and they make a negligible contribution to heat transport [110]. Thus, the accumulation of polar optical phonons would result in the localization of a significant portion of the thermal energy near the hot spot until the optical phonons decay to faster moving acoustic phonons (about 10 ps), which carry the energy away from the hot spot [112]. Moreover, this accumulation of optical phonon modes near the hot spot can build up

over time, causing in such way the electrons to experience an increased number of scattering events, with negative impact on electron transport [111, 113]. However, thermal simulations, accounting for the local phonon population and dynamics, were not carried out in the present study, but are part of our present on-going effort and will be reported in a future work.

6.6 Feedback Capacitance and Output Impedance

A drawback of gate-connected field plates is the increased feedback capacitance due to the stronger coupling between drain and gate. This feedback capacitance is related to the drain-to-gate coupling occurring through the air-gap and through the semiconductor. In particular, the coupling through the air-gap is enhanced by the reduced contact distance and by a medium with large dielectric constant. On the other hand, the coupling through the semiconductor is reduced by an extended depletion region at the drain side. This latter aspect was further investigated, and the electron density profile with different L_{arm} values is shown in Figure 6.14. As we can see, there is a large difference in terms of depletion region between the different L_{arm} values in the SiN case. As we can see, the depletion region extension basically corresponds to L_{arm}. Moreover, the depletion region extension is greatly increased for a high drain voltage of $V_D = 18\ V$.

We performed small-signal analysis in order to obtain C_{gd} from the imaginary part of the Y_{12} parameter, and quantify the overall effect of the air-gap and semiconductor drain-to-gate coupling. A simplified estimation of C_{gd} can be obtained from the basic FET small-signal model reported in Section 2.3. The dependence of the imaginary part of Y_{12} on the frequency, for several different Fourier Transform windows, is shown in Figure 6.15 for the device biased at $V_D = 6\ V$ in the $L_{arm} = 500\ nm$ configuration. As we can see, the spreading due to the different Fourier window lengths is contained, and Y_{12} exhibits a fairly linear behavior over the range of frequencies of [0-100] GHz. This is important because it indicates that, in this frequency range, the model represents a good approximation, and C_{gd}, computed as the slope of the linear

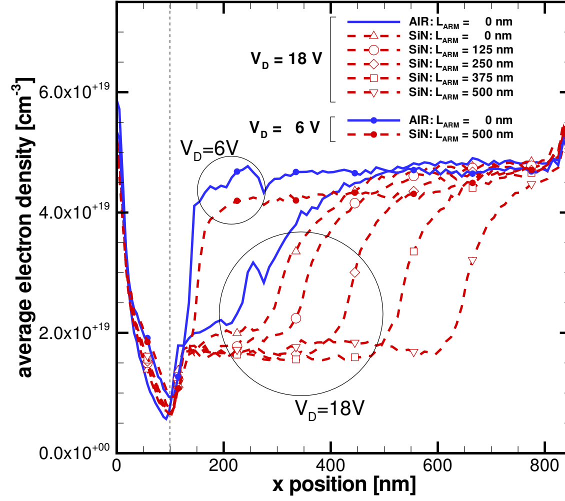

Figure 6.14: Electron density profile along the channel versus different T-gate arm lengths (L_{arm}) in the air (solid line) and SiN (dashed line) cases with a T-gate height of 80 nm. The devices are biased for maximum f_T at $V_G = -2\ V$ and $V_D = 6\ V$ (filled symbols) and maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$ (hollow symbols). The vertical dashed lines represent the physical gate length drain-side end. The gate source-end is set as origin of the axis.

function, has constant value. Moreover, C_{gd} obtained from low-frequency experimental measurements can be extrapolated and assumed to be a good estimation for higher frequencies in this range. Observing the two different set of curves related to the air and SiN cases, we can see how the latter configuration exhibits a much larger C_{gd} (87 fF/mm versus 254 fF/mm). This is mainly due to the increased air-gap coupling caused by the larger SiN dielectric constant. The depletion region plays a minor role in this comparison, because the difference between the two configurations in terms of depletion region is limited at such low drain bias of 6 V as shown before. That is not the case at larger bias such as $V_D = 18\ V$.

The C_{gd} values related to the different gate geometries and biasing values were extracted at a frequency of 40 GHz, corresponding to a typical GaN HEMTs application frequency range located in the Ka-band, and are shown in Figure 6.16. As we can see, C_{gd} increases relatively linearly with L_{arm} for $V_D = 6$ in both the SiN and air

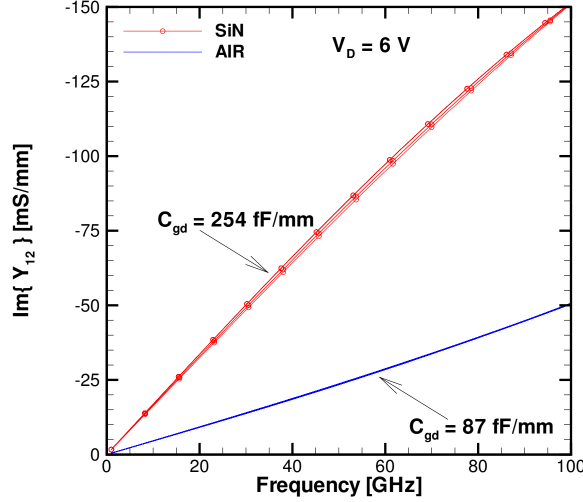

Figure 6.15: Imaginary part of Y_{12} versus frequency for the air (solid line) and *SiN* (line with circles) case with $L_{arm} = 500\ nm$ and T-gate height of 80 *nm*. Several extraction Fourier Transform window lengths are overlaid in the plot. The devices are biased for maximum f_T at $V_G = -2\ V$ and $V_D = 6\ V$. The *SiN* case exhibits a larger negative slope than the air one, corresponding to a larger C_{gd} of 254 fF/mm versus 87 fF/mm.

configurations. Moreover, the combined effect of *SiN* and T-gate extension greatly increase C_{gd} raising its value from 60 fF/mm in the case of air without T-gate to 254 fF/mm in the *SiN* with T-gate one. On the other hand, when observing the C_{gd} behavior at $V_D = 18\ V$ we first notice how C_{gd} is much smaller than when $V_D = 6\ V$, due to the more extended depletion region. This explains why f_{max}, heavily affected by the feedback capacitance, achieves its maximum experimental value at a larger drain bias than f_T. Furthermore, at such drain voltage, C_{gd} is not particularly affected by L_{arm}, and the difference between the air and *SiN* cases C_{gd} is small. All these effects can be explained in terms of an increased depletion region extension, which is due to the larger dielectric in the *SiN* configuration that enhances the field plate action as discussed before. Therefore, the increased air-gap capacitive coupling between drain and gate is associated with a reduced coupling through the semiconductor when the device with *SiN* is biased with a large drain bias. Indeed, a larger T-gate height (130 *nm* instead 80 *nm*) results in a lower C_{gd} at $V_D = 6\ V$, but in a larger C_{gd} at $V_D = 18\ V$ due to reduced

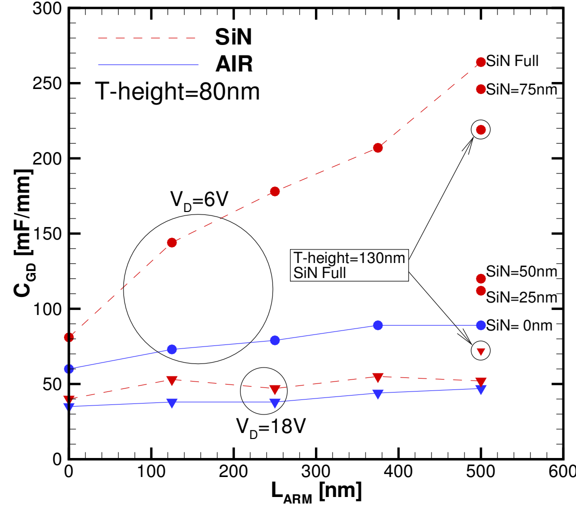

Figure 6.16: Gate-to-drain feedback capacitance, extracted at 40 GHz through small-signal analysis from $Im\{Y_{12}\}$, in air (solid line) and SiN (dashed) cases with increasing L_{arm} and T-gate height of 80 nm. The devices are biased at $V_D = 6\ V$ (circles) and $V_D = 18\ V$ (triangles) with $V_G = -2\ V$ in both cases.

ability of the field plate to deplete the gate-drain region. Concerning the SiN thickness, with a layer of 25, 50, and 75 nm the difference in terms of C_{gd} is, respectively, the 12%, 18%, and 90% of the difference between the SiN (*i.e.* 80 nm) and the air cases with $L_{arm} = 500\ nm$ and $V_D = 6\ V$ as also shown in Figure 6.16.

The effect on the large-signal output impedance at 25 GHz of this drain-voltage-dependent variation of C_{gd} is shown in Figure 6.17. The output characteristic is obtained by applying a sinusoidal voltage signal at the drain contact, while the gate voltage is kept constant at the bias value. The resulting output drain current is then plot versus the applied sinusoidal drain voltage and overlaid to the $I_d - V_d$ static characteristic. This large-signal dynamic characterization fully exploits the ability of our Full-Band Monte Carlo to accurately model the highly non-linear, out-of-equilibrium transport mechanisms due to high-energy carriers inside the devices. As we can see in the SiN case with $L_{arm} = 500\ nm$, the phase difference between output current and output voltage is significantly increased at low drain voltages. This large variation of the

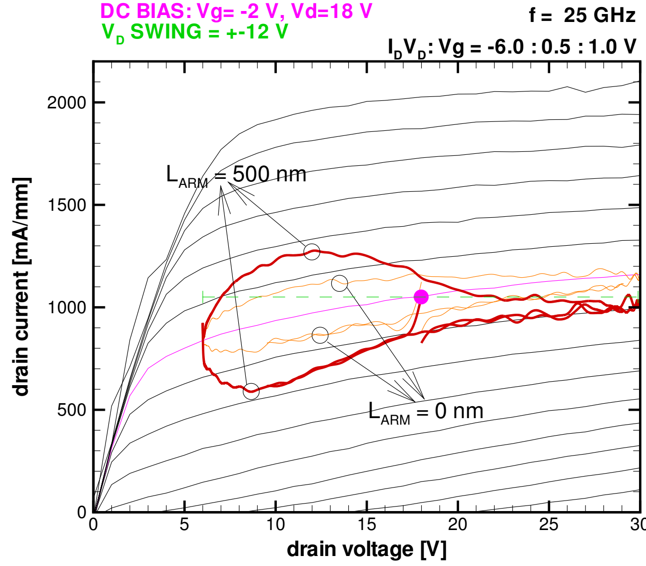

Figure 6.17: Impact of L_{arm} on the large-signal output impedance at 25 GHz in a device with SiN dielectric constant and a T-gate height of 80 nm. The output characteristic is obtained by applying a sinusoidal voltage signal at the drain contact, while the gate voltage is kept constant at the bias value. The device is biased for maximum f_{max} at $V_G = -2\ V$ and $V_D = 18\ V$.

large-signal output impedance during the device operations can negatively affect the design of the output matching network and also the device linearity in power amplifiers as discussed in Section 2.5.

Once C_{gd} is know, we can extract the drain-to-source capacitance, C_{ds}, from the imaginary part of Y_{22} as also reported in Section 2.3. However, since Y_{22} is largely dominated by C_{gd}, the extraction of C_{ds} resulted troublesome. In particular, within numerical error, we could not assess any evidence of a dependence of C_{ds} on the passivation layer dielectric constant and L_{arm}. We obtained $C_{ds} = 13\ fF/mm$ at $V_D = 6\ V$, and $C_{ds} = 8\ fF/mm$ at $V_D = 18\ V$. On the other hand, the strong dependence of f_T, also extracted from the same set of small-signal analysis, on L_{arm} in the SiN case is confirmed by the obtained values of $f_T = 100, 86$, and 75 GHz associated to, respectively, $L_{arm} = 250, 375$, and 500 nm at $V_D = 18\ V$. This dependence is due to the second gate control region introduced by the field plate and the related longer total transit time.

By combining the previous results obtained from our DC and RF analysis, we computed the total gate capacitance, C_g, and the source-to-gate capacitance, C_{gs}, from the values of f_T, g_m, and C_{gd} as described in Eq.(2.22). In particular, by evaluating C_g at a bias point of $V_G = -2\ V$, $V_D = 6\ V$ ($f_T = 150\ GHz$, $g_m = 370\ mS/mm$, and $C_{gd} = 60\ fF/mm$), we obtained $C_{gs} = 332\ fF/mm$. On the other hand, by evaluating C_g at a bias point of $V_G = -2\ V$, $V_D = 18\ V$ ($f_T = 135\ GHz$, $g_m = 340\ mS/mm$, and $C_{gd} = 35\ fF/mm$), we got $C_{gs} = 365\ fF/mm$ which is only about 10% larger than the $V_D = 6\ V$ case. This is due to the little influence of the applied drain voltage on the source side of the device and on the source-to-gate coupling associated to the extension of the source-side depletion region

6.7 Conclusions

In this chapter, we showed that the dielectric constant of the passivation layer has a large impact on the nano-scale carrier dynamics due to the change in the device electrostatics. Thus, the dielectric constant of the passivation affects key quantities such as carrier distribution and transit time, effective gate length, phonon emission/absorption, intrinsic capacitances, and also plays a major role in affecting the RF performance of deep sub-micron devices. The main point is that there is in fact a trade-off between the beneficial effects of a surface passivation layer (from reduction of surface states) and a degrading effect due to the change in electrostatics arising from the increased dielectric constant. Therefore, this work represents a valuable tool for device engineers and researchers to better understand how the performance of *GaN* HEMTs is affected by nano-scale phenomena.

Chapter 7

Circuit-Device Simulations

7.1 Introduction

The Monte Carlo technique allows a detailed description of the nanoscale phenomena occurring in a device. However, the circuit packaging and the extrinsic parasitic elements due to interconnections can play a role in the experimental characterization of an actual device as important as the intrinsic device behavior. In particular, simulating only the intrinsic device may lead to an overestimation of DC and RF performance. More specifically, the reactive parasitic elements related to interconnections can dominate the ultra-fast device operation at high frequency, and their effect may no longer be neglected. Furthermore, the interaction between the device and the external circuitry, such as matching networks, load impedance and transmission lines, must included with either an analytical or a non-analytical approach, in order to properly reproduce the device operating conditions as well as to obtain a more complete RF characterization [16, 17]. In this chapter, a description is provided of the different network solvers implemented in our code to incorporate the effect of external network circuit elements (which are referred here to as the ”extrinsic network”) into the device directly simulated by the CMC code (called here the ”intrisic device”). The complete system made by the intrinsic device plus the extrinsic network is referred to as the ”extrinsic device”.

It must be underlined that the implemented network solvers do not have the capability of solving an arbitrary circuit topology, but they rather include pre-determined standard topologies. This choice is due to several reasons. First, the parasitic elements, relevant for microwave device simulations, are usually related to relatively simple topologies (*i.e.* series or shunt elements as discussed in Section 2.3). Moreover, the effect of complex topologies relevant for RF circuits, for instance matching networks, can be often coupled with the device simulations with simplified approaches that do not

need for directly solving the whole topology, as it will be clear in Section 7.4 and Chapter 8. Furthermore, the scope of physical Circuit-Device simulations is not to reproduce the whole circuit behavior, but to rather focus on the few circuit elements nearby the device that are relevant for the proper reproduction of the device working conditions. In particular, the simulation of the whole circuit would be prohibitive if terms of computational burden, which represents the downsides of a particle-based device simulation techniques such as the Monte Carlo one. On the other hand, compact models, whose parameters can be extracted either from experimental or simulated devices (see Section 2.3), allow including the behavior of a device in a time-efficient manner in actual circuit simulator tools. Therefore, the implementation of a complicated user interface able to provide the solver with a general topology, as well as the implementation of the numerical methods required for the automatic resolution of the associated system of equations, is beyond the scope of this work. However, future extensions of the code implementing the network solvers described in this chapter could include such functionalities.

7.2 AC Small-signal Network Solver

This network solver analytically includes the effect of an extrinsic network in post-processing. Its main application is the evaluation of the effects of extrinsic parasitic elements, described in Section 2.3, on the device RF small-signal performance by embedding such parasitics in the small-signal characterization. The main asset of this solver is that there is no need for running new simulations if different circuit elements are to be included in the device characterization

Once the intrinsic device is characterized as a two-port network in terms of Y- or Z-parameters (see Chapter 2), the effect of an extrinsic network can be included, within the small-signal analysis framework, by simply considering the Y- or Z-parameters of the network itself. These extrinsic network parameters are then combined with the

intrinsic device ones to obtain the extrinsic device parameters:

$$\mathbf{Y}_{ext} = \mathbf{Y}_{int} + \mathbf{Y}_{net} \tag{7.1}$$

or

$$\mathbf{Z}_{ext} = \mathbf{Z}_{int} + \mathbf{Z}_{net}. \tag{7.2}$$

Choosing which one of two approaches is to be used depends on the specific topology of the extrinsic network. In general, series extrinsic network elements are more conveniently included by using the Z-parameters, while shunt elements are modeled by using the Y-parameters.

This AC small-signal network solver can be also used in combination with the FDTD network solver described later in Section 7.3. In such case, the AC small-signal solver analyzes the extrinsic current obtained as output of the CMC (intrinsic device simulator) coupled with the FDTD network solver (extrinsic network simulator). However, this approach is not recommended due to the large noise introduced by the FDTD network solver and its limitation in terms of high-frequency components.

In case of FET devices, the effect of a series resistor and a series inductor associated to the gate, source, and drain contacts can be taken into account by using the approach reported in [53], and shown in Figure 7.1. The network Z-parameters are:

$$\mathbf{Z}_{net} = \begin{bmatrix} (R_g + R_s) + j\omega(L_g + L_s) & R_s + j\omega L_s \\ R_s + j\omega L_s & (R_d + R_s) + j\omega(L_d + L_s) \end{bmatrix}. \tag{7.3}$$

Even though extrinsic network elements have values which are usually independent of the device width, their effect changes if the width of the simulated device is scaled. In case the simulated device has a different width than the real device - as it usually happens -, the network elements to be included can be scaled according to the real width in order to have the same effect they would have on the real device. For

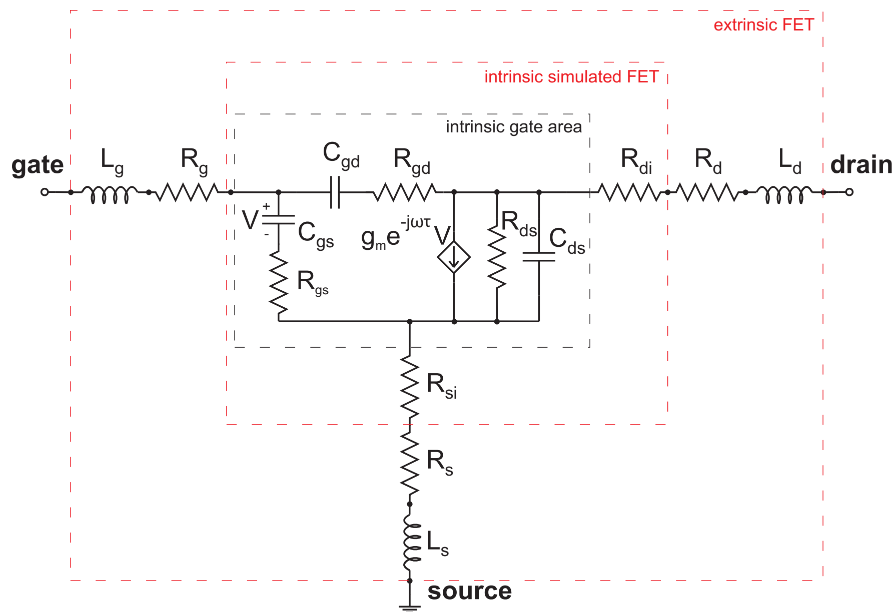

Figure 7.1: Small-signal equivalent circuit model of the intrinsic device and the extrinsic parasitic network model. R_{si} and R_{di} are, respectively, the source access resistance and the drain access resistance associated to the part of semiconductor between the source or drain contact and the gate.

instance, the effect of a series resistor of 1 Ω with a device of 100 μm width is the same of a series resistor of 100 Ω with a device of 1 μm width (*i.e.* same voltage drop). This because the 100 μm width device has a current 100 times larger than the 1 μm one.

It must be noted that in a real device R_s and R_d change the DC operating point of the intrinsic device due to the DC voltage drop associated to these series resistances. However, due to the fact that the inclusion of these extrinsic elements is performed in post-processing, the resulting extrinsic device Y- and Z-parameters (which are bias dependent) do not reflect this change in the DC operating point. Therefore, the intrinsic device parameters should be extracted at a bias that includes the voltage drop associated to the series resistances included in post-processing. On the other hand, there is not in general a DC voltage drop associated to the gate resistance, R_g, due to the fact that the gate current usually has only an AC component. However, this is not the case if gate leakage (due for example to tunneling) is included in the simulation, resulting in a gate current with a DC component.

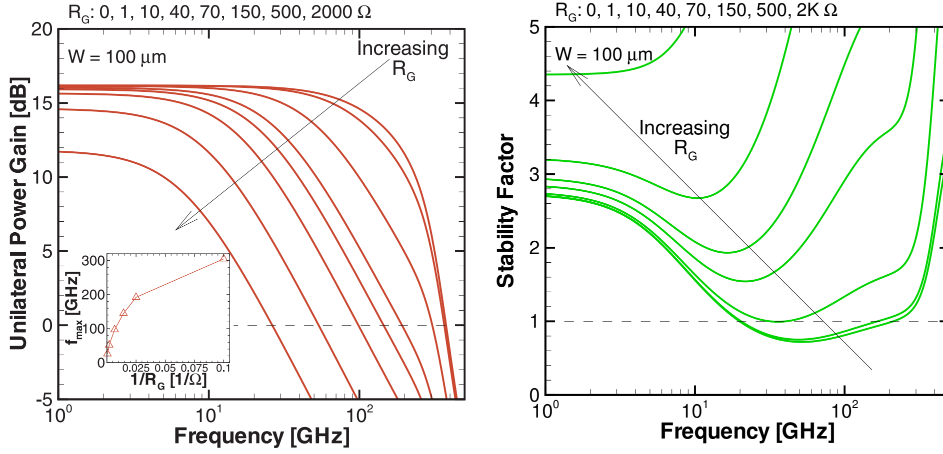

Figure 7.2: Effect of the gate resistance, R_G, on the Unilater Power Gain (left) and Stability Factor (right). f_{max} versus R_G^{-1} is also reported in the small inset in the left figure.

In Figure 7.2, we can see an application of the AC small-signal circuit solver where the gate resistance, R_G, was embedded in the AC small-signal characterization obtained through the Cellular Monte Carlo device simulator. As we can see, the Unilateral Transducer Power Gain is heavily reduced by the increasing R_G, which in turn results in a lower f_{max}. On the other hand, we can verify that the Stability Factor is progressively pushed away from the region of non-unconditional stability of the device (*i.e* stability factor < 1) due to the R_G resistive dumping stabilizing effect (*i.e.* lossy stabilization network). This stability improvement, as well as the power gain drop, is especially evident at high-frequency (*i.e.* above 10 GHz in this example) due to the fact that the higher is the frequency the higher is the displacement current flowing through the gate, which results in a larger impact of the gate resistance due to the related larger input voltage drop. Moreover, by observing f_{max} versus R_G^{-1}, as reported in the small inset in the left side of Figure 7.2, we can see that this relation is sub-linear with a behavior roughly corresponding to $f_{max} \sim \sqrt{R_G^{-1}}$, which is the common analytical dependence between f_{max} and R_G expected in FETs [56] (see Eq.(2.23)).

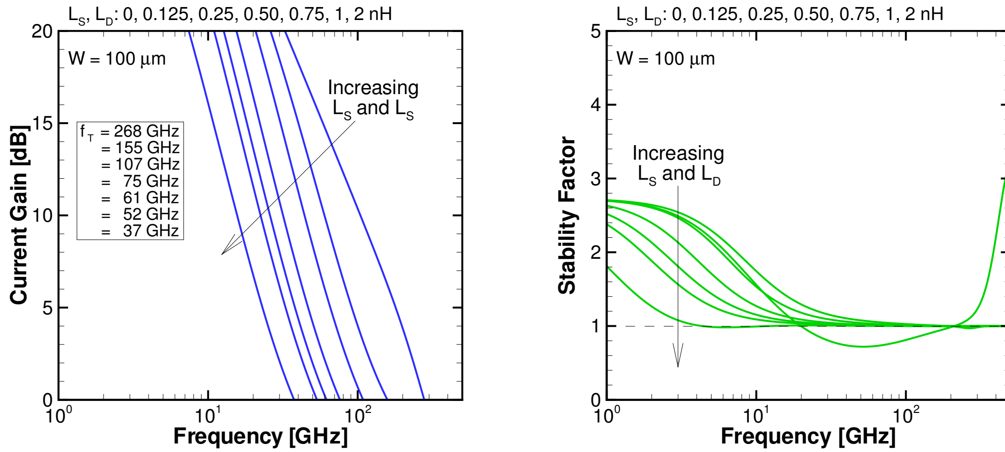

Figure 7.3: Effect of the source and drain inductance, L_S and L_D, on the Short-circuit Current Gain (left)and the Stability Factor (right).

In Figure 7.3, we can also see the effect of the source and drain parasitic inductances (*i.e.* inductive coupling due to wire bonding), L_S and L_D , on the small-signal RF performance of the device. As we can see, the Short-Circuit Current Gain is negatively affected by these parasitic inductances, resulting in a lower f_T. Moreover, we can verify that the stability factor is negatively affected as well. In particular, the larger the parasitic inductances, the more the stability factor is pushed closer to the region of non-unconditional stability of the device (*i.e* stability factor < 1). This is mainly due to the source inductor that introduces a positive feedback, which occurs at lower frequencies as the inductance value is increased as suggested by plot in the right side of Figure 7.3.

7.3 Finite-Difference Time-Domain (FDTD) Network Solver

We have implemented a real-time self-consistently coupled Finite-Difference Time-Domain (FDTD) circuit solver that has the capability of solving simple external networks of lumped circuit elements connect to the contacts. This is done by self-consistently coupling the intrinsic device simulated by the CMC with the time-domain external network solver:

- Intrinsic device simulator (CMC) input/output:
  - **input**: intrinsic voltages as given by the network solver.
  - **output**: intrinsic currents supplied to the network solver.
- FDTD Network solver input/output:
  - **input**:
    - extrinsic voltages: potential values provided by the user.
    - intrinsic currents: given by CMC intrinsic device simulator.
  - **output**:
    - intrinsic voltages: to be applied to the intrinsic device contacts.
    - extrinsic currents

Within this approach, the voltage values specified by the user are not directly applied to the intrinsic device contacts, but are modified by the extrinsic network (*i.e* voltage drop). Thus, the external network is self-consistently coupled to the intrinsic device by using the intrinsic currents obtained by the CMC simulation to solve the network in the time-domain and to calculate the intrinsic voltages that are applied to the intrinsic device contacts. Thus, we can characterize the behavior of the intrinsic device combined with the extrinsic network (*i.e.* ”extrinsic device”). This self-consistent procedure is performed in real-time during the simulation for each single time step, T_s. Moreover, since the network is solved in the time-domain through a finite difference time domain approach, there are no limiting assumptions regarding the type of input signals (*i.e.* extrinsic voltage waveforms), the device operation regime (*i.e.* DC, transient small-signal, transient large-signal, linear, non-linear, single-tone, multi-tone).

The greatest advantage of a time-domain solver is the ability to handle very strong non-linearities (see chapter 2) and transient phenomena. In fact, as long as the

non-linearities are continuous, the time resolution (*i.e.* the simulation time step) can always be increased enough so that the circuit voltages and currents change very little from one step to the other. In particular, in our Monte Carlo device simulations, the time step corresponds to the Poisson solver time step as discussed in Section 1.1, which is usually on the order of 10^{-15} seconds. Thus, we can properly simulate devices with very high-frequency input as well as strong non-linearities due to this extremely high time resolution.

The drawback related to this approach is that the FDTD network solver performs a real-time signal processing (smoothing and down-sampling, see Section 7.3 and 7.3) of the intrinsic currents driving the network in order to improve the signal-to-noise ratio and to improve the solver stability. The adjustable parameters related to these signal-processing algorithms limit the maximum value of the harmonic frequency components present in the signal. The optimum parameters depends on the network topology, the contact to which the network is connected, the amplitude of the signals and their derivatives, the signal frequency, and whether there is the generation of harmonics.

FDTD Network Topologies

As discussed in the Introduction, pre-determined standard topologies have been implemented in the code rather than realizing a solver capable to process a general topology provided by the user. In particular, the FDTD equations directly provide the relevant currents and voltages in the circuit have been implemented for each different topology. However, in case of a future extension of the code, a suitable iterative method for automatically obtain the solution of a system of linear FDTD equations generated in the process of solving a general topology, is the Gauss-Seidel Method [114] or similar relaxation methods. The FDTD network topologies available in the current implementation are described in the following sub-sections.

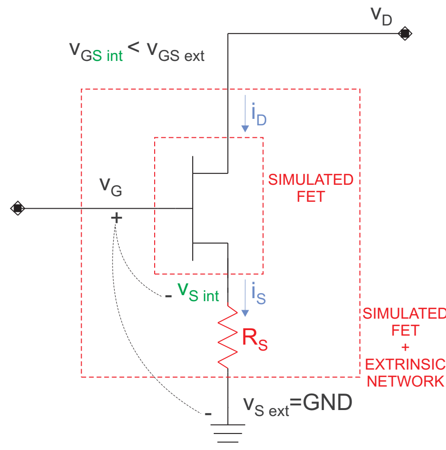

Figure 7.4: Example of the series impedance topology: FET with a series source resistance. The FET is simulated through the CMC device simulator, while the extrinsic network is included by a FDTD solver self-consistently coupled with the device simulator.

SERIES IMPEDANCE

This topology simulates a resistor, R, and an inductor, L, both connected in series to the specified contact. As series capacitor is not allowed within this topology because it would block the DC-biasing of the intrinsic contact. The difference between extrinsic and intrinsic voltages is due to the sum of the resistor and the inductor voltage drops. An example of a series resistor, R_S, connected to the source contact of a FET can be seen in Figure 7.4. In such topology, v_{GSint} seen by the intrinsic device is smaller than v_{GSext} applied to the extrinsic device (*i.e.* specified in the simulation file) due to the voltage drop across R_S. The FDTD equations associated to this topology, with respect to a case where the series impedance is connected to the source contact, are the following:

$$\Delta v_R(nT_s) = i_S(nT_s) \cdot R_s, \tag{7.4}$$

$$\Delta v_L(nT_s) = \frac{[i_S(nT_s) - i_S(nT_s - T_s)] \cdot L}{T_s}, \tag{7.5}$$

$$\Delta v_{tot}(nT_s) = \Delta v_{R_s}(nT_s) + \Delta v_{L_s}(nT_s), \tag{7.6}$$

$$\Delta v_{Sint}(nT_s) = \Delta v_{tot}(nT_s) + v_{Sext}(nT_s), \tag{7.7}$$

where the series inductor, L, is not included in this example.

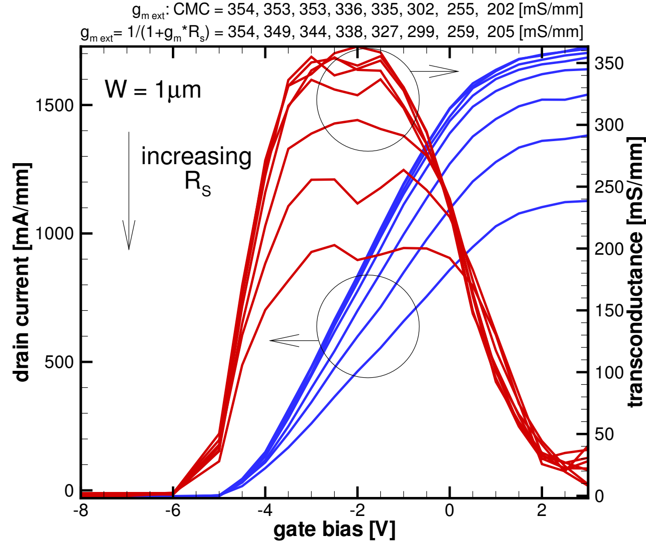

Figure 7.5: Example of the series impedance topology: resulting DC extrinsic drain current and transconductance versus gate voltage. The peak *gm* values obtained through the FDTD/CMC simulation (higher top line) and the analytical ones obtained through Eq(7.8) (lower top line) are also reported for comparison.

By observing the simulation results in Figure 7.5, we can see that this reduced v_{GSint} results in a extrinsic device with lower drain current and transconductance. The simulated results agrees with the well-know analytical expression [66]:

$$g_{m_{EXT}} = \frac{g_{m_{INT}}}{1 + g_{m_{INT}} \cdot R_S}. \tag{7.8}$$

Moreover, the behavior of v_{Sint} when a sinusoidal AC signal is applied to the gate can be see in Figure 7.6, where both the DC and AC value of v_{Sint} are affected by R_s.

The effect of a drain resistor, R_D, is shown in Figure 7.7. In this topology, v_{Dint} is smaller than the actual drain voltage applied to the extrinsic device, v_{Dext}. Thus, a larger drain voltage has to be applied such that the intrinsic device actually operates with an intrinsic drain voltage equal to the one that there would be if R_D were zero. This results in a shift of the extrinsic device saturation drain voltage, V_{Dsat}, and an ”expansion” of the $I_D - V_D$ characteristics.

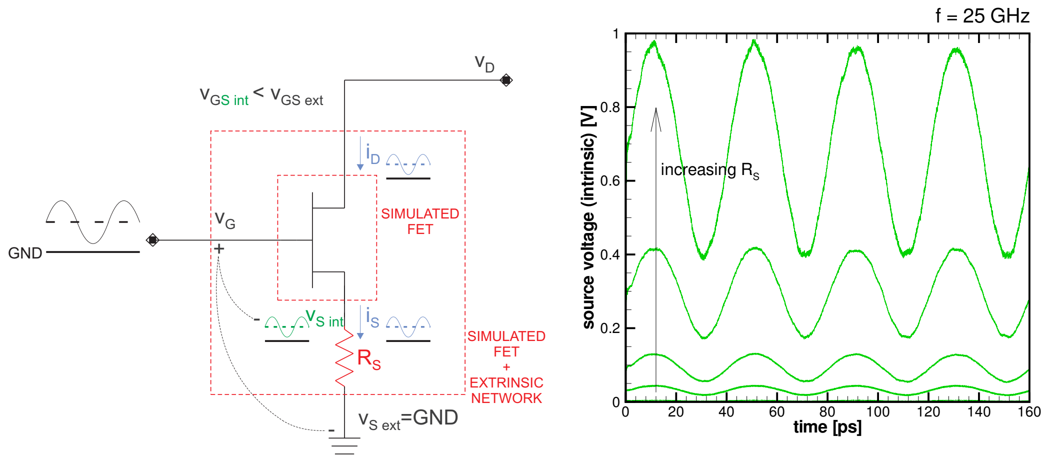

Figure 7.6: Example of the series impedance topology: resulting AC intrinsic source voltage.

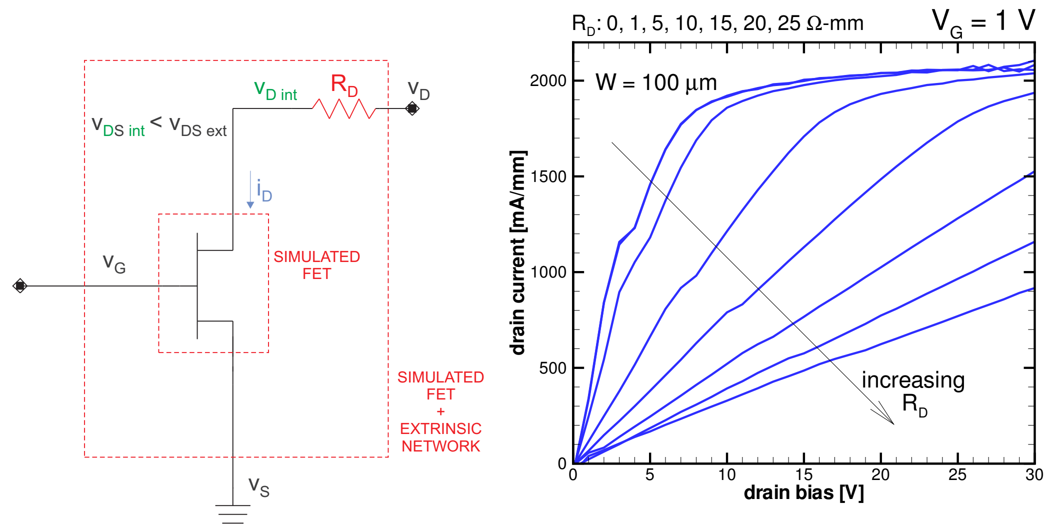

Figure 7.7: Example of the series impedance topology: FET with a series drain resistor (left) and resulting drain current versus drain voltage (right).

An example with a series reactive elements such as inductor, L_S, connected to the source contact of a FET can be seen in Figure 7.8. In such topology, v_{GSint} seen by the intrinsic device corresponds to the derivative of the source current, i_S. As we can see, the resulting v_{Sint} has correctly a zero DC component and is 90 degrees out of phase with respect to i_S. The large transient in the initial part of the voltage waveform is due the time required by the FDTD/CMC solver to achieve a stable numerical solution

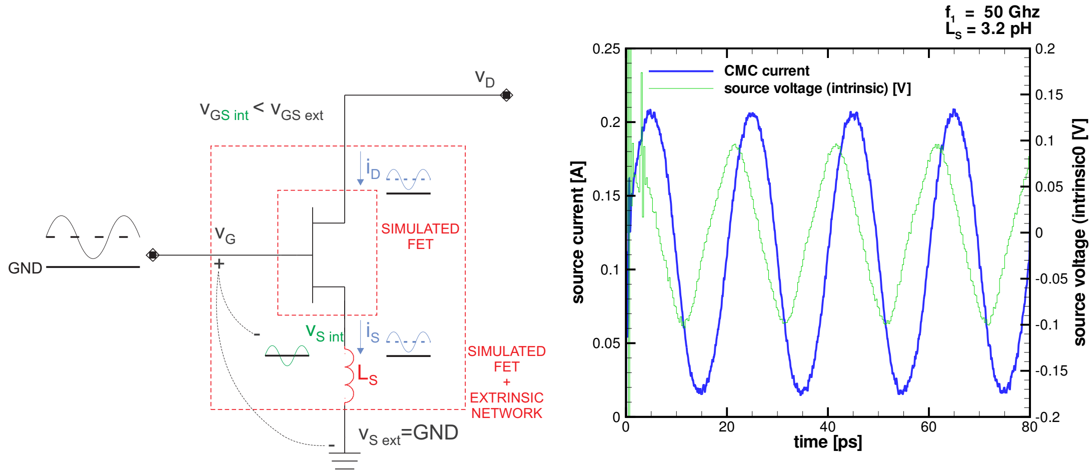

Figure 7.8: Example of the series impedance topology: FET with a series source inductor (left) and resulting AC intrinsic source voltage versus time (right).

because the initial currents and voltages (which correspond to a case without external network) do not satisfy the network solution.

SHUNT IMPEDANCE

This topology simulates a resistor, R, and a capacitor, C, both connected in shunt to the specified contact. A shunt inductor is not allowed within this topology because it would short DC-wise the contact to the ground. With respect to a case where the shunt impedance is connected to the drain contact, the difference between intrinsic (*i.e.* i_{Dint}) and extrinsic current (*i.e.* i_{Dext}) is the difference between the intrinsic current and the sum (*i.e.* i_{shunt}) of the currents flowing through the resistor (*i.e.* i_R) and the capacitor (*i.e.* i_C). The FDTD equations associated to this topology, with respect to a case where the shunt impedance is connected to the drain contact, are:

$$i_R(nT_s) = -v_D(nT_s)/R, \tag{7.9}$$

$$i_C(nT_s) = -\frac{[v_D(nT_s) - v_D(nT_s - T_s)] \cdot C}{T_s}, \tag{7.10}$$

$$i_{shunt}(nT_s) = i_R(nT_s) + i_C(nT_s), \tag{7.11}$$

$$i_{Dext}(nT_s) = i_{Dint}(nT_s) - i_{shunt}(nT_s). \tag{7.12}$$

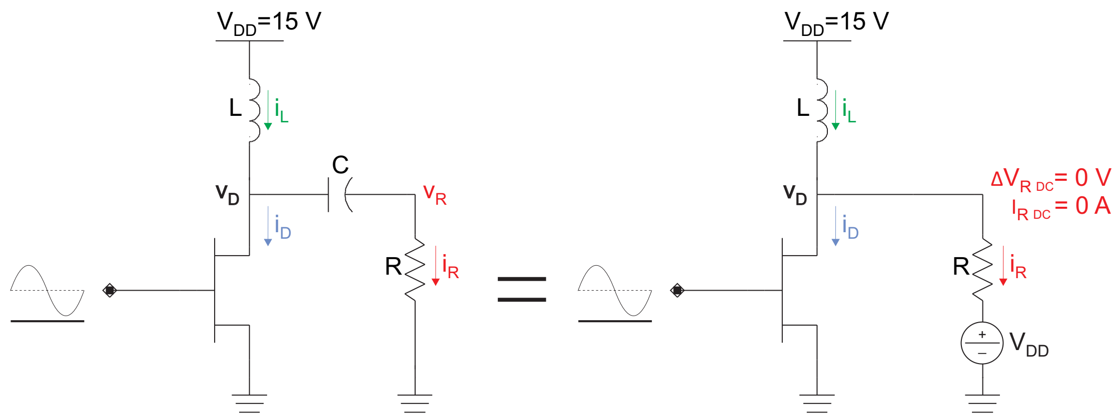

Figure 7.9: Example of the bias tee topology: resistive load connected to the drain contact.

BIAS TEE

This model simulates a bias tee topology with a resistive load, R, connected to the specified contact. An example of such network topology connected to the drain contact can be seen in Figure 7.9. According to this topology, if both the capacitor, C, and the inductor, L, are large enough, they respectively act as a blocking capacitor (*i.e.* AC short and DC open: no DC current through C) and a RF choke inductor (*i.e.* AC open and DC short: no AC current through L). In such case, the inductor current, i_L, will have zero AC-component, while the AC-component of the resistor current, i_R, will be equal to the AC-component of drain current ,i_D, (*i.e.* all the transistor output current delivered to the load). On the other hand, i_R always has zero DC-component (due to the presence of the capacitor) regardless the value of C, while the device contact voltage, v_D, DC-component is always equal to V_{DD} (due to the presence of the inductor) regardless the value of L. Based on this last observation, the blocking capacitor does not need to be explicitly simulated because we can force i_R to have a zero DC-component by connecting a DC-generator (set to V_{DD}) between R and the ground. In such way, the DC-voltage drop across R is zero, and zero has to be the DC-component of i_R.

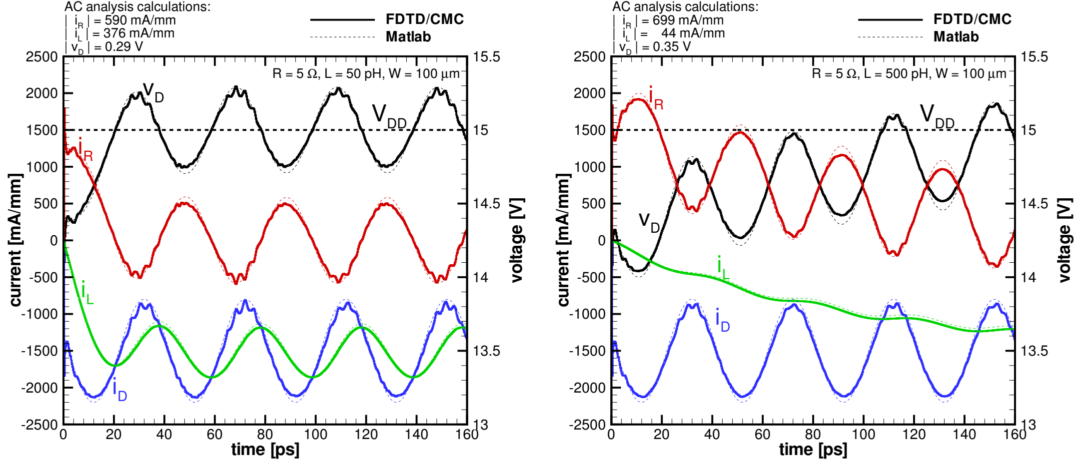

Figure 7.10: Example of the bias tee topology: output currents and voltages with an input gate voltage sine of 25 GHz. The device has width of 100 μm. LEFT: $R = 5\ \Omega$, $L = 50\ pH$. RIGHT: $R = 5\ \Omega$, $L = 500\ pH$.

The FDTD equations associated to this topology, with respect to a case where the bias tee is connected to the drain contact, are:

$$i_L(nT_s) = \frac{i_D(nT_s) \cdot R + [i_L(nT_s - T_s) \cdot L/T_s]}{R + L/T_s}, \tag{7.13}$$

$$i_R(nT_s) = i_L(nT_s) - i_D(nT_s), \tag{7.14}$$

$$v_D(nT_s) = V_{DD} - i_R(nT_s) \cdot R. \tag{7.15}$$

Observing Figure 7.10, we can see that, as we increase the value of L, we effectively reduce the AC component of i_L. On the other hand, the time constant of the circuit is increased as well ($\tau = L/R$), requiring in such way a longer simulation time before exhausting the transient time and achieving a steady-state situation ($i.e.$ $i_{L_{AC}} = 0$, $i_{R_{AC}} = i_{D_{AC}}$, $i_{L_{DC}} = i_{D_{DC}}$, $i_{R_{DC}} = 0$, and $v_{D_{DC}} = V_{DD}$). In the same plots, a comparison is also reported with the solution obtained through *Matlab* by applying the same set of FDTD equations to a sinusoidal current. We can verify a good agreement between the waveforms obtained with the two method as well as the fact that the actual FDTD/CMC output waveforms are slightly compressed, with respect to the pure sinusoidal *Matlab* waveforms. This has to be expected, and is due to the device non-linear behavior cap-

tured by the CMC simulations. Such behavior is due to the large input voltage signal as well as the self-consistent coupling between FDTD and CMC where i_D is affected by v_D, calculated by FDTD, that is fed back to the CMC as input. The values obtained from AC analysis analytical hand calculations are also reported in the same figure, confirming the agreement between AC theory and simulations.

Real-time Smoothing Algorithm

The FDTD network solver performs a real-time smoothing of the intrinsic currents driving the network in order to improve the signal-to-noise ratio and to improve the solver stability. The real-time smoothing algorithm is essentially a FIR filter [115] that processes the most recent and the previous samples of the intrinsic current at the end of each time step to produce a smooth current. This procedure effectively filters-off high-frequency noise.

$$y(nT) = \sum_{m=0}^{L-1} b_k \cdot x(nT - mT), \tag{7.16}$$

where T is the Poisson time-step in the CMC code, L is the length of the filter, b_k are the filter coefficients, and $x(nT)$ is a sample at the time step nT.

In order to preserve the shape of the noisy input signal, the filter must satisfy the non-distortion conditions stating that the magnitude of the filter frequency response must be constant and the phase of the filter frequency response must be linear:

$$|H(\omega)| = A, \tag{7.17}$$

$$\angle H(\omega) = C \cdot \omega, \tag{7.18}$$

where A and C are two constants.

With such a filter, the relative amplitudes of the frequency components is preserved as well as their relative time delays (*i.e.* no time-dispersion of the frequency components is introduced by the filter).

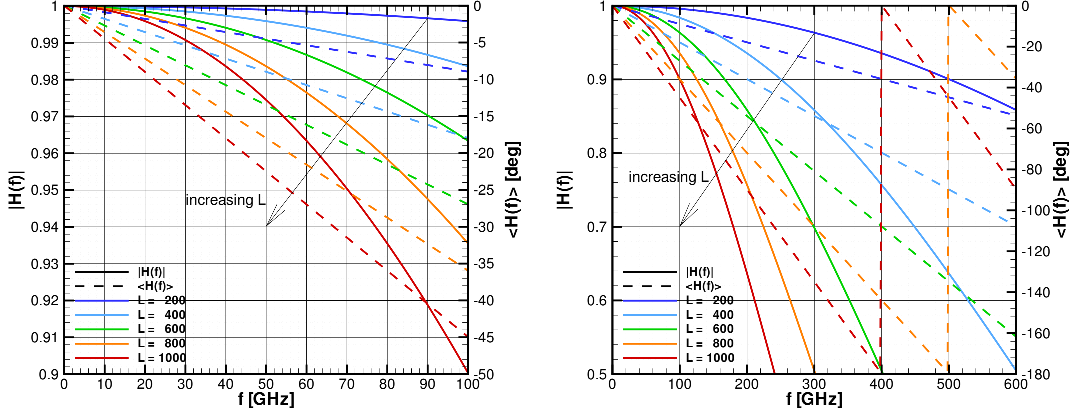

Figure 7.11: Moving average smoothing frequency response function between [0-100] GHz (left) and [0-600] GHz (right).

In our case, we want to have have a smooth version of the instantaneous intrinsic current to be used as driving current to solve the network. Thus, we want to preserve the shape and absolute amplitude (*i.e.* $|H(\omega)| = 1$ and $\angle H(\omega) = C \cdot \omega$). Moreover, we want to avoid any time delay between the instantaneous current and its smoothed version (*i.e.* $\angle H(\omega) = 0$).

The implemented smoothing algorithm is a real-time moving average. The smoothed current sample value is given by all the past instantaneous current values stored in a buffer of increasing length l:

$$i_{smooth}(nT) = \sum_{m=0}^{l-1} \frac{1}{l} \cdot i(nT - mT). \tag{7.19}$$

Therefore, the smoothing algorithm starts to produce an output without waiting for the buffer to be full. Once the buffer is full (*i.e.* $l = L$, where L is the moving average window), a faster update equation is used:

$$i_{smooth}(nT) = i_{smooth}(nT - T) - \frac{i(nT - LT)}{L} + \frac{i(nT)}{L}. \tag{7.20}$$

The frequency response of this moving average filter for different window length L is shown in Figure 7.11. The low-pass nature of the filter is evident as well as the

linear phase response. The low-pass behavior (which is required to filter-off the noise) can negatively affect the relative amplitude of the frequency components of interest in the output signals. This is because the output high-frequency components can have a smaller amplitude than the input high-frequency components resulting in the linear distortion of the signal. On the other hand, the linear phase guarantees that the relative time-delay of the frequency components is not modified, avoiding dispersion and further linear distortion. However, the fact that the filter phase is not zero results in a delay between filter input and output (*i.e.* delay between intrinsic current and its smoothed version).

The duration L of the moving average window in the smoothing filter impacts the stability of the network solver by changing the signal-to-noise ratio of the driving currents. The driving currents are the smoothed version of the intrinsic currents, and they are used in combination with the extrinsic voltages by the network solver to solve the external networks. The amplitude attenuation and phase delay is related to L, and the higher the frequency of the signal the higher is this attenuation and delay. There is therefore a trade-off between the driving current signal-to-noise ratio (improved by a longer smoothing filter) and the difference between the actual intrinsic current and the resulting driving current (worsen by a longer smoothing filter). The duration L of the moving average window can be guessed by using Figure 7.11.

Down-sampling

The fundamental time step associated to current and potential waveforms in the CMC code is the Poisson time-step, T. However, the FDTD network solver can operate using a larger time-step, $T_s = NT$, obtained by down-sampling, in order to improve the signal-to-noise ratio of the derivatives of currents and voltages required to solve networks with reactive elements as shown in Figure 7.12.

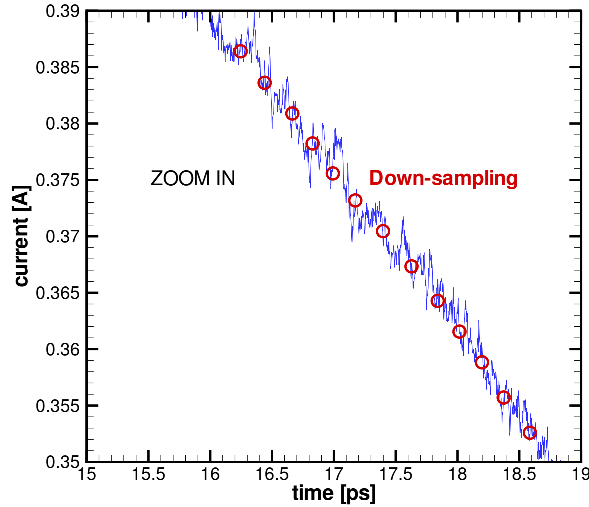

Figure 7.12: Down-sampling example.

The network is solved only at the down-sampled time steps, so the smoothed intrinsic voltages and extrinsic currents, calculated by the network solver, are down-sampled signals with a sample-and-hold approach [115] between samples. The driving currents are also a down-sampled version of the smoothed intrinsic current signal generated by the smoothing algorithm. However, the smoothing algorithm is synchronized to the Poisson time-step T. This means that it processes the samples and produces and output at the end of each T regardless the down-sampling ratio used by the FDTD network solver to represent the signals and to solve the network. The down-sampling ratio, N, limits the maximum frequency components present in the waveform spectrum, so it should be chosen properly in order to have down-sampled a driving current that resembles the original intrinsic current. Even when no reactive elements are included, the low-pass nature of the down-sampling procedure allows improvement of the stability of the network solver by removing high-frequency noise.

7.4 Harmonic Balance Frequency-Domain Network Solver

This is an iterative, self-consistently coupled, frequency-domain network solver that has the capability of solving external networks of lumped circuit elements connect to the contacts. This solver can be extended to also include the effect of distributed circuits that are easily described in the frequency-domain (*i.e.* transmission lines).

The Harmonic Balance algorithm (HB) should be mainly used with AC simulations to overcome the following limitations of the previously described network solvers:

- **FDTD** (see Section 7.3): Solving networks with reactive elements with large values by using time-domain techniques requires long simulation time due to the related long transient times (*i.e.* RC, L/C time constants). An example of this issue can be see in in Figure 7.10, where the larger is the reactive element (*i.e.* L) the longer is the time required by the transient part of the response to decay. So, time-consuming Monte Carlo particle-based simulators are somewhat poorly suited for this kind of analysis. Moreover, the FDTD algorithm introduces a lot of noise in the simulator output because it directly calculates the intrinsic voltages starting from the noisy intrinsic currents produced by the CMC code. The real-time smoothing algorithm used to reduce this noise introduces a phase delay in the solution (see Section 7.3). Furthermore, a large number of reactive elements can result in rather complicated and large finite difference equations. This has a further detrimental effect on the numerical noise introduced by the FDTD solver.

- **AC small-signal** (see Section 7.2): Small-signal modeling assumes that the AC signal amplitude is much smaller than the DC value, and that the system response is linear. The small-signal parameters allow predicting device performance and behavior on the base of analytical equations. However, non-linearity arises as soon as the amplitude of the AC component of the signal cannot be considered

small. Therefore, the small-signal AC analysis fails to analytically predict large-signal performance and the impact of an external network (see Section 2.5).

The Key Idea of the Harmonic Balance Algorithm

Harmonic Balance exploits the fact that the steady-state behavior of a network of linear passive elements can be more easily obtained by a frequency domain analysis rather than by a time-domain one. For instance, the steady-state output of a *RC* low-pass filter with sinusoidal input can be readily obtained in the frequency-domain by multiplying the input signal phasor by the complex frequency response of the filter. On the other hand, a time-domain approach implies solving a differential equation, which results in an exponential time-domain solution involving a transient time (*i.e.* the RC time constant). Only once the transient response has decayed, the filter output will corresponds to the steady-state solution. Moreover, the numerical solution of the differential equation itself can be quite problematic in case of complex networks.

The key idea of the Harmonic Balance algorithm is to self-consistently couple a time-domain solution of the active non-linear device, provided by the Monte Carlo particle-based device simulator, with a frequency domain solution of the extrinsic network. This is practically accomplished by simply substituting a network element, or a network of many elements, with sinusoidal voltage generators connected to the corresponding contact. Such generators, whose amplitude and phase are obtained from the frequency domain solution, will directly generate the network steady-state time-domain solution. In such way, the effect of the impedance associated to a network element, or the overall impedance of a network of many elements, is coupled to the intrinsic device simulator by emulating its effect in terms of the steady-state voltage waveforms seen by the intrinsic device. Most of the extrinsic network elements can be often assumed to be passive and linear. This is the case of lumped circuit elements as well as distributed elements such as transmission lines. On the other hand, the intrinsic devices simulated by the Monte Carlo code are generally active and non-linear (*i.e.* transistors or diodes).

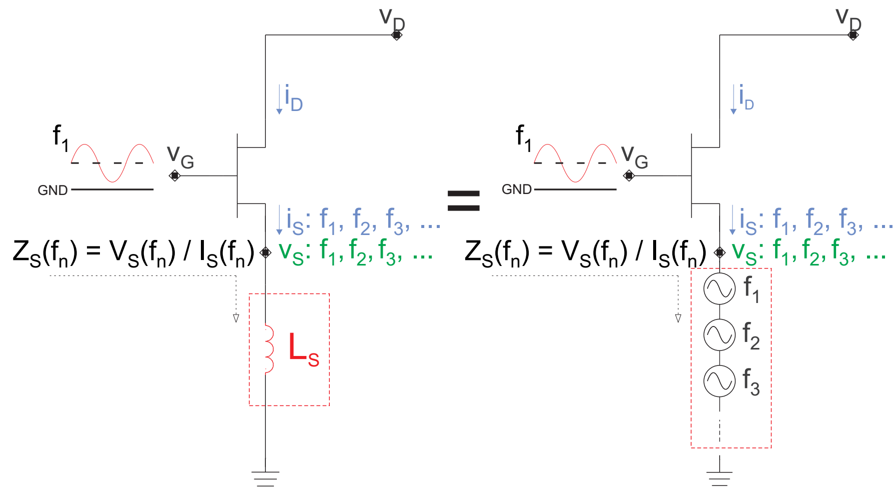

Figure 7.13: Schematic example of the general HB algorithm: the source inductor is emulated by generating the voltage sinusoids at the different harmonics.

The frequency at which the sinusoidal voltage generators are tuned depends on the frequency components of the current waveforms flowing through the contact to which the extrinsic network is connected. In particular, the device input waveform frequency components (*i.e.* gate voltage in a FET) will determine the frequency components of the resulting output current waveforms. However, in case of large-signal simulations (which are the main purpose of the HB network solver), the higher order harmonics as well as the intermodulation products due to non-linearity (see Section 2.5) should be also taken into account by adding voltage generators tuned at the corresponding frequency. Therefore, the accuracy of HB (*i.e.* the ability to correctly emulate a certain impedance by reproducing the corresponding voltage waveform) is related to the number of frequency components included in the solutions. In general, the farther the current waveform is from a sinusoid, the large is the number of harmonic components required to properly emulate the resulting voltage waveform.

The example in Figure 7.13 can better explain the Harmonic Balance algorithm idea. The simulation results, obtained by using the algorithm described in Section 7.4 to model the effect of the source inductor of the example are shown in Figure 7.14 obtained for a device biased for Class-A and Class-B operation (see Section 2.5). In the

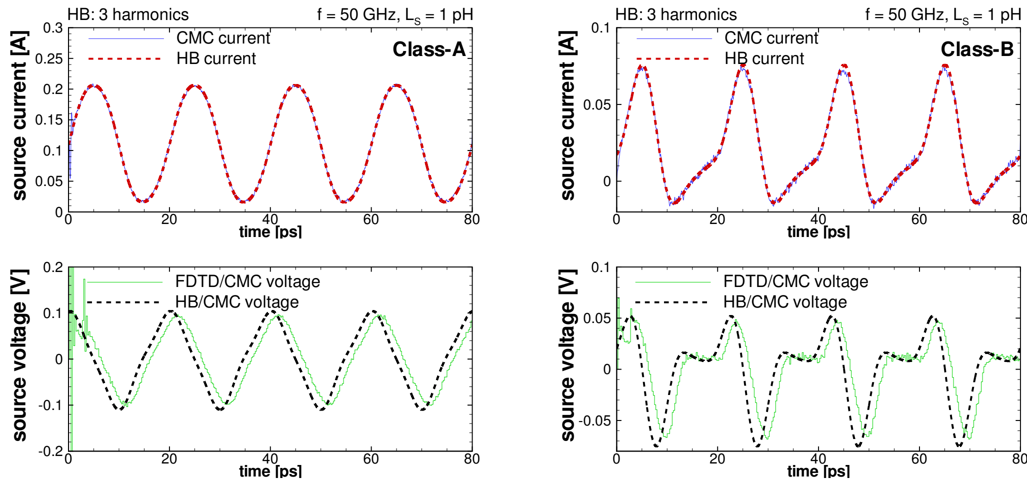

Figure 7.14: Simulation results of a source inductor emulated through HB by generating the voltage sinusoids for the first three harmonics. The device is biased for Class-A (left) and Class-B (right) operations. The FDTD solver solution is also showed for comparison. The CMC output current has been smoothed in post-processing by the plotting program only for graphic purposes. The FDTD solver solution delay is due to the real-time moving average filter.

Class-A case, reported in the left side of Figure 7.14, the source current mainly contains the fundamental harmonic plus some weaker higher harmonics resulting from the distortion introduced by the large-signal operations. In particular, the third harmonic is commonly related to the compression of the peak value of the sinusoidal output current. Therefore, up to three harmonics are included by HB in this set of simulations. The FDTD network solver described in Section 7.3 is also used for comparison. However, the FDTD solution exhibits a small delay due to the real-time moving average filter discussed in Section 7.3. As we can see, there is a 90 degree shift between source current and source voltage obtained by HB as already depicted in Figure 7.8 for the FDTD case. Moreover, we can see that the calculated source voltage, in both the HB and FDTD cases, has a triangular shape rather than sinusoidal. This confirms that both the solvers obtained a similar solution and that the device output waveforms contain harmonics beyond the fundamental as expected. Moreover, the agreement between HB and FDTD confirms that including the first three harmonics in HB is enough to achieve a satisfactory solution in this case.

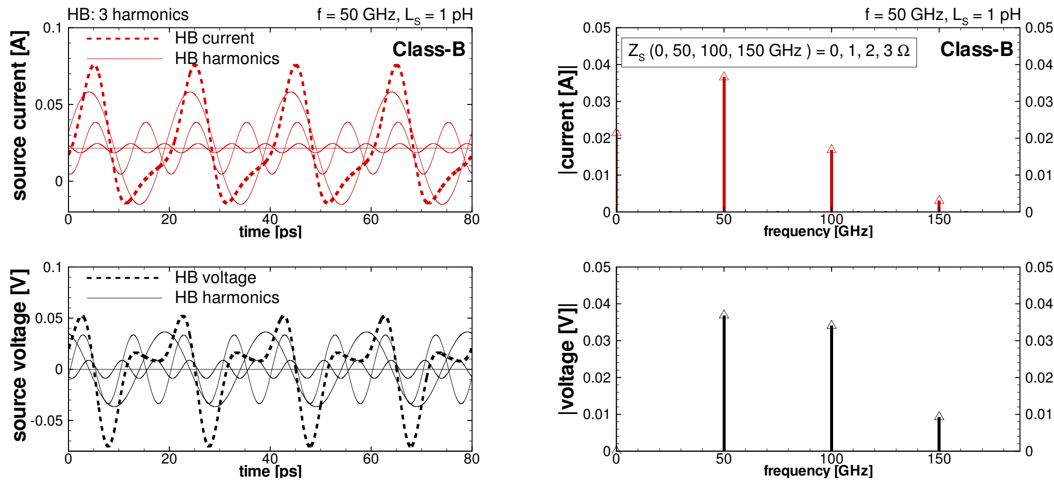

Figure 7.15: Harmonic content of the analyzed source current, and of the voltage waveform generated by the Harmonic Balance algorithm.

On the other hand, including only one harmonics in the HB solution would result in a purely sinusoidal source voltage.

We can verify that also in the Class-B case reported in the right side of Figure 7.14 both HB and FDTD produce similar results, and that including the first three harmonics is enough to achieve a satisfactory solution in this case. This is further confirmed by the fact that the current used by HB to perform its calculations, and extracted by considering the first three harmonics of the Fourier transform of the CMC output current, agrees well with the CMC output current as shown in the top panel of the right side of Figure 7.14. In the left side of Figure 7.15, we can also actually see how the superposition of the voltage sinusoids, generated by the voltage generators tuned at the specified harmonic frequencies, is able to accurately reproduce the device output voltage waveforms. Moreover, from the right side of Figure 7.15, we can see that the source current of the device biased in Class-B has as expected a strong second harmonic component. Furthermore, the source inductor, associated to an increasing source impedance with increasing frequency, produces a source voltage waveform containing even more relevant higher harmonics.

In general, Harmonic Balance can be used to couple time-domain device simulations with circuits that are difficult, and computationally expensive, to analyze in the time-domain such as high-Q matching networks (see Section 2.5 and Section 8), transmission lines, RF filters, and multi-port sub-networks. All of these structures are described by the Y-,Z-, or S-parameters such that the frequency-domain approach of Harmonic Balance can be easily and efficiently applied.

The General Harmonic Balance Algorithm

The Harmonic Balance analysis is applicable to a wide variety of problems related to microwave circuits such as power amplifiers, frequency multipliers, and mixers. Due to the fact that the Harmonic Balance approach directly calculates the circuit steady-state response, it works particularly well when a circuit has a mix of long and short time constants. In fact, it was originally proposed to solve the problems inherent in analyzing such circuits [116].

In Figure 7.16(a), a simple circuit is reported to explain in a general way the Harmonic Balance algorithm. As we can see, an RF signal, $v_s(t)$, consisting of a DC component and a sinusoidal component, is connect to a diode. The $v_s(t)$ waveform does not need to be a pure sinusoid, but it must be periodic. For the sake of simplicity, we focus here on the single tone case. The diode excited by a large RF input signal generates additional harmonics due to its non-linear behavior. The time-domain exponential relation between voltage and current waveforms, which is at the origin of the diode non-linearities especially when driven with large-signals, allows us to easily calculate the diode current once the diode voltage is known. On the other hand, the impedance connected to the diode is generally linear, but exhibits a different value for each of the different frequency components, and its behavior is consequently not easily described in the time-domain. This difficulty can however be overcome by first assuming a certain spectral composition of the voltage, $V(\omega_i)$. One then creates an equivalent linear circuit, as shown in Figure 7.16(b), which includes only the linear components of the

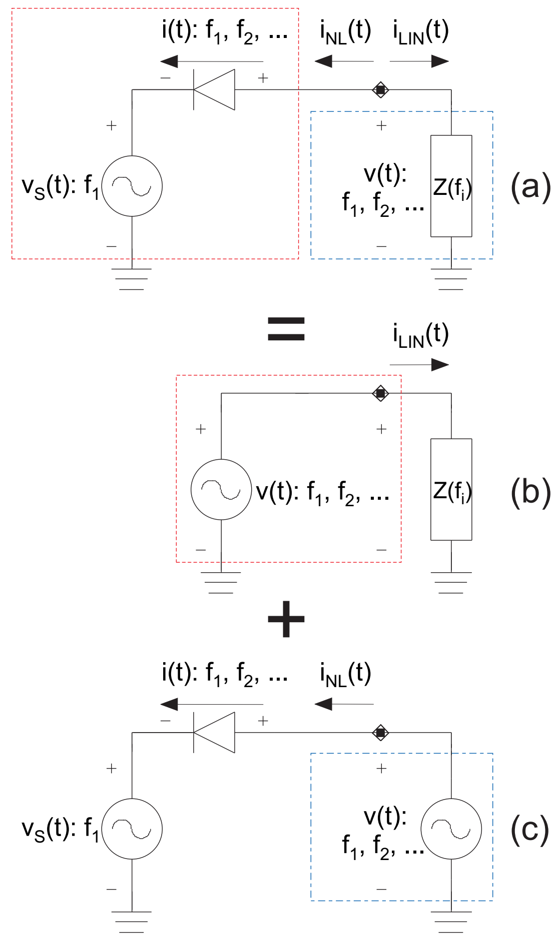

Figure 7.16: The circuit including a diode excited by a large RF signal (a) can be divided into a pair of equivalent circuits, one describing the linear part (b), and another the non-linear part (c).

circuit (*i.e.* the impedance) that can be analyzed easily in the frequency domain, giving:

$$I_{LIN}(\omega_i) = \frac{V(\omega_i)}{Z(\omega_i)}. \tag{7.21}$$

Then, through the the Inverse-Fourier transform of $V(\omega_i)$ we can obtain the voltage waveform $v(t)$:

$$\mathscr{F}^{-1}\{V(\omega_i)\} \longrightarrow v(t). \tag{7.22}$$

Next, we create an equivalent non-linear circuit, shown in Figure 7.16(c), which includes only the non-linear components of the circuit (*i.e.* the diode), and can be better solved in the time-domain. Thus, we can find algebraically the current waveform in the

diode, $i_{NL}(t)$, from $v(t)$:

$$i_{NL}(t) = I_{sat} e^{\left[\frac{v(t)-v_S(t)}{V_T}-1\right]}, \tag{7.23}$$

where I_{sat} is the diode reverse bias saturation current and V_T is the thermal voltage. The only issue so far is that we do not know whether the initial voltage harmonic composition actually represents a solution of the circuit. However, through the the Fourier transform of $i_{NL}(t)$ we can obtain the harmonic composition of $I_{NL}(\omega_i)$:

$$\mathscr{F}\{i_{NL}(t)\} \longrightarrow I_{NL}(\omega_i). \tag{7.24}$$

At this point, we can verify whether the Kirchoff's current law is simultaneously satisfied for each frequency component in the circuit of Figure 7.16(a):

$$I_{LIN}(\omega_i) + I_{NL}(\omega_i) = 0. \tag{7.25}$$

If Eq.(7.25) is satisfied, then $V(\omega_i)$ we is a solution. Otherwise, another $V(\omega_i)$ has to be chosen and verified.

The general solution process is summarized below:

1. Create an initial estimate of $V(\omega_i)$ for each ith frequency component of interest. This initial guess can be zero for every ω_i.

2. Calculate $I_{LIN}(\omega_i)$ from $V(\omega_i)$ by using Eq.(7.21).

3. Obtain $v(t)$ from the Inverse-Fourier transform of $V(\omega_i)$.

4. Calculate $i_{NL}(t)$ from $v(t)$ by using Eq.(7.23).

5. Obtain $I_{NL}(\omega_i)$ from the Fourier transform of $i_{NL}(t)$.

6. Substitute $I_{LIN}(\omega_i)$ and $I_{NL}(\omega_i)$ in Eq.(7.25). This equation will not be probably satisfied, so define an error function for each ith frequency component:

$$I_{LIN}(\omega_i) + I_{NL}(\omega_i) = I_{ERR}(\omega_i). \tag{7.26}$$

Note that each $I_{ERR}(\omega_i)$ is implicitly a function of *all* the other voltage components $V(\omega_i)$ due to the non-linear response of the diode (*i.e.* non-linear output current).

7. Evaluate whether $|I_{ERR}(\omega_i)|$ is small, according to a certain metric, for each ith frequency component of interest *simultaneously.* In the negative case, use an appropriate numeric method to modify $V(\omega_i)$, such that $|I_{ERR}(\omega_i)|$ will be reduced at the end of the next iteration, and repeat the process from step 2.

In general, more complex circuits can have a large number of both linear and nonlinear circuit elements. These elements can be grouped two sub-circuits, one linear and the other nonlinear. The linear sub-circuit can be treated as a multi-port network and directly described by its Y-,Z-, or S-parameters, or by some other multi-port matrix in the frequency domain. On the other hand, the nonlinear elements are modeled by their global I/V or Q/V characteristics in the time-domain, which fully take into account any non-linear behavior, and the corresponding multi-port matrix is obtained through subsequent Fourier analysis. The combination of the matrix of the linear and non-linear sub-circuits through the Kirchoff's laws, can result in multiple sets of non-linear equations whose zeros can be found by using appropriate numerical methods such as the Newton's Method. More details about solving complex networks through the Harmonic Balance algorithm and the numerical techniques required to solve the related system of equations can be found in the Maas book [61].

The Self-consistent Harmonic Balance / Monte Carlo Algorithm

As stated at the beginning of this chapter, we did not implement solvers able to solve a general circuit topology, but we rather focused on some pre-set network models relevant for transistor simulations. This is also the case for our implementation of the Harmonic Balance algorithm self-consistently coupled with the Monte Carlo device simulator. So, we do not have to worry about automatically solving sets of non-linear

equations describing a general topology. Moreover, we focus our iterative approach on the comparison between a desired target impedance and the impedance resulting from the chosen voltage waveform, while the conventional approach, as previously described, focuses on the discrepancy between non-linear and linear currents resulting from the chosen voltage waveform. The difference between the two approach lies in the fact that with our approach we let change the value of the linear circuit elements in order to satisfy the Kirchoff's laws even when the wrong voltage waveform is chosen as solution. On the other hand, the conventional approach focuses on the discrepancy between the non-linear current, which represent the response of the non-linear device output to the chosen voltage waveform, and the linear-current, which represents the response to the chosen voltage waveform of the defined linear elements (*i.e.* the defined impedance). With this latter approach, the Kirchoff's laws are not satisfied if the right voltage waveform is not chosen, but the linear circuit elements are not allowed to change value. In both cases, the algorithm has to be iterated until the desired impedance is achieved (our approach), or the non-linear and linear currents are equal (conventional approach). The two methods are *completely equivalent* and they just represent a different interpretation of the results produced by the Harmonic Balance iterative procedure.

Within our approach, we start from an initial combination of voltage sinusoids connected to the contact, to which the extrinsic network is associated, to run the first simulation. The solution of the network is represented by a certain combination of voltage sinusoids such that the complex ratio between the voltage and current frequency components (expressed as phasors) corresponds to the equivalent impedance of the network at each frequency of interest. Therefore, the initial combination of voltage sines represents an initial estimate of the solution, and corresponds to the steps 1 and 3 of the conventional approach.

Then, the actual synthesized extrinsic network equivalent impedance is determined in post-processing at the end of the first simulation through Fourier transform:

$$
\begin{aligned}
\widetilde{Z}(\omega_i) &= |Z(\omega_i)| e^{j\angle Z(\omega_1)} \\
&= \frac{\widetilde{v}(\omega_i)}{\widetilde{i}(\omega_i)} = \frac{|v(\omega_i)|}{|i(\omega_i)|} e^{j(\angle v(\omega_i) - \angle i(\omega_i))},
\end{aligned}
\tag{7.27}
$$

where $\widetilde{v}(\omega_i)$ and $\widetilde{i}(\omega_i)$ are the complex phasors of the drain current and voltage sinusoids at the ith frequency component ω_i. This correspond to the step 2 of the conventional approach.

The magnitude and phase of the complex impedance are adjusted in order to match the target impedance by changing magnitude and phase of the voltage generator. Then, a new simulation is run where the updated values are obtained as follow in our implementation:

$$
|v'(\omega_i)| = |v(\omega_i)| \left(\frac{|Z_{target}(\omega_i)|}{|Z(\omega_i)|} \right), \tag{7.28a}
$$

$$
\angle v'(\omega_i) = \angle v(\omega_i) + (\angle Z_{target}(\omega_i) - \angle Z(\omega_i)), \tag{7.28b}
$$

where $\widetilde{Z}_{target}(\omega_i)$ is the desired impedance at the ith frequency component ω_i. This correspond to the step 7 of the conventional approach.

Due to the fact that the HB solver is self-consistently coupled to the Monte Carlo device simulator, the change in the voltage stimulus will obviously result in a different device behavior in terms of output current. So, even after the correction of the voltage sinusoids, the desired impedance will not likely be achieved. For this reason, the simulation and the subsequent impedance analysis must be iterated until the actual emulated impedance is close enough, according to some predefined metric, to the desired impedance for each ith frequency component ω_i. In our implementation the following metric is used to assess the convergence:

$$
\begin{cases}
|Re\{Z_{target}(\omega_i)\} - Re\{Z(\omega_i)\}| & \leq |Re\{Z_{err_{MAX}}(\omega_i)\}| \\
|Im\{Z_{target}(\omega_i)\} - Im\{Z(\omega_i)\}| & \leq |Im\{Z_{err_{MAX}}(\omega_i)\}|,
\end{cases}
\tag{7.29}
$$

where $Z_{err_{MAX}}(\omega_i)$ is the defined maximum impedance error corresponding to the ith frequency component ω_i. This is the analogous of the step 6 and 7 of the conventional approach.

Note that the adjusting procedure based on through Eq.(7.28a) and (7.28b) at the end of each cycle uses a linear correction approach. Therefore, as general rule of thumb, the more non-linear are the device operations, the more difficult (*i.e.* more iterations, slow convergence, or even divergence) is going to be achieving the convergence of the algorithm with such implementation. A possible non-linear correction approach could exploit the knowledge of the harmonic content of the output current, and use it to evaluate the non-linearity of the device from the amplitude of the generated harmonics. At that point, this information could be used to calculate a parameter of non-linearity, which combined with a modified versions of Eq.(7.28a) and (7.28b), would allow to apply a non-linear correction. It must be noted that the convergence must be *simultaneously* achieved for each ith frequency component ω_i. Thus, the correction must be continuously applied to *every* i^{th} frequency component ω_i as long as this condition is not met. This because, there is interaction between the frequency components in a non-linear system (see Section 2.5).

Windowing Effect and Single Iteration Length

The duration of the device simulation (*i.e.* the length of each single HB/CMC iteration) can heavily affect the accuracy of the output current harmonic content extraction. In particular, simulating a certain periodic waveform with a finite number of periods corresponds to the windowing (*i.e.* truncation) of the same waveform infinitely extended. In such case, the Fourier transform of the signal waveform corresponds to the convolution of the *sinc* function with the discrete spectrum of the periodic infinitely extended waveform. This results in a spectrum where there is a *sinc* function superimposed to each discrete frequency component. The width of each *sinc* function main lobe directly affects the ”resolution” of the Discrete Fourier Transform, DFT, which represents the

ability of resolving two adjacent frequency components. This width is ultimately inversely proportional to the number of simulated periods.

The number of HB iterations needed to obtain a solution is generally not know (especially when large non-linearities result in a slow convergence), thus the number of simulated periods plays a critical role for time-consuming particle-based Monte Carlo device simulations. Thus, there is stringent trade-off between accuracy and total simulation time in self-consistently coupled Harmonic Balance / Monte Carlo simulations. This trade-off also depends on the type of simulation (*i.e.* single tone, multi tone) due to the number of harmonics and intermodulation products involved (see Section 2.5).

Convergence Improvement under Relaxation

When an iterative algorithm like Harmonic Balance is combined with the Monte Carlo technique, issues such as low signal-to-noise ratio, linear corrections applied to highly non-linear devices, and discrepancies between measured and actual current harmonic content of the device can result in poor convergence. In particular, a low signal-to-noise ratio often occurs when the amplitude of the AC signal is relatively small as well as when trying to include higher harmonics in the solution, which are usually just a fraction of the fundamental frequency. Moreover, a linear correction approach, where for instance the voltage amplitude is simply doubled each time the target impedance is twice the emulate impedance at the fundamental frequency, likely experiences convergence issues when large non-linearities result in a non-linear increase of the fundamental frequency amplitude due to harmonic generations. Furthermore, the windowing problem, related to the Fourier transform of the harmonic content of the output current, can heavily affect the proper extraction of the amplitude and phase of closely spaced frequency components. In such cases, the iterative algorithm can be affected by slow convergence, divergence, or oscillations around the solution. The cause of that is in many cases an over-correction of amplitude and phase of the new voltage sines, that are applied at the next iteration, due to the issues discussed above.

In order to avoid over-corrections, a dumping factor can be used to reduce the amplitude and phase correction when a convergence issue is detected (*i.e.* increasing discrepancy between the emulated and desired impedance in subsequent iterations). In general, the real part of the impedance is indirectly related to the amplitude of the generated voltage sines. On the other hand, the imaginary part of the impedance is indirectly related to the phase of the generated voltage sines. Moreover, in order to improve the flexibility of this convergence improvement approach, separate dumping factors are used for both amplitude and phase as well as for each different frequency component.

In order to properly implement a reduced correction step approach, the amplitude correction has to be converted from a multiplicative form to an additive form:

$$|v'(\omega_i)| = |v(\omega_i)| + \Delta v(\omega_i)/k_{Re}(i), \tag{7.30}$$

$$\Delta v'(\omega_i) = |v(\omega_i)| \left(\frac{|Z_{target}(\omega_i)|}{|Z(\omega_i)|} \right) - |v(\omega_i)|, \tag{7.31}$$

where $k_{Re}(i)$ is a positive integer number, and represents the dumping factor associated to the ith frequency component ω_i.

The phase correction is already in additive form:

$$\angle v'(\omega_i) = \angle v(\omega_i) + \Delta\angle v(\omega_i)/k_{Im}(i), \tag{7.32}$$

$$\Delta\angle v(\omega_i) = \angle Z_{target}(\omega_i) - \angle Z(\omega_i), \tag{7.33}$$

where $k_{Im}(i)$ is a positive integer number, and represents the dumping factor associated to the ith frequency component ω_i.

The dumping factor, $k_{Re}(i)$ or $k_{Im}(i)$, associated to a certain i^{th} frequency component ω_i is progressively increased every time a convergence issue is detected according to:

$$\begin{cases} Re\{Z_{err}(\omega_i)_N\} \geq Re\{Z_{err}(\omega_i)_{N-1}\} \\ \Longrightarrow k_{Re}(i)_N = k_{Re}(i)_{N-1} + 1, \end{cases} \tag{7.34}$$

$$\begin{cases} Im\{Z_{err}(\omega_i)_N\} \geq Im\{Z_{err}(\omega_i)_{N-1}\} \\ \Longrightarrow k_{Im}(i)_N = k_{Im}(i)_{N-1} + 1, \end{cases} \tag{7.35}$$

where N is the number of present iterations, and $Z_{err}(\omega_i)$ is:

$$Z_{err}(\omega_i) = |Z_{target}(\omega_i) - Z(\omega_i)|. \tag{7.36}$$

On the other hand, the dumping factor, $k_{Re}(i)$ or $k_{Im}(i)$, associated to a certain i^{th} frequency component ω_i is progressively decreased if the opposite condition is met as long as the dumping factor is not already equal to the minimum value of 1.

Difference between the HB Solver and the Other Network Solvers

We conclude the description of the Harmonic Balance approach with a schematic discussion of the main differences between the HB algorithm and the previously described network solvers:

- **FDTD** (see Section 7.3): This is a self-consistent, real-time, solver. Each time step the CMC output intrinsic output is used to solve the extrinsic network (real-time) and to calculate the intrinsic voltages to be applied to the intrinsic device that in turn will modify the CMC output intrinsic current of the next time-step (self-consistency). FDTD allows to analyze the steady-state behavior of a long enough AC simulation only after the transient response has decayed. FDTD introduces a large amount of noise through the noisy intrinsic voltages resulting from the network calculation. It also introduces a phase delay and an attenuation of the higher frequency components in the solution due to the smoothing average filter. On the other hand, a time-domain solutions intrinsically includes a large number of frequency components only limited by the fundamental time step. In the present CMC code, the FDTD solver output discriminates between intrinsic and extrinsic currents/voltages.

HB is a self-consistent, non-real-time, steady-state iterative solver. In other words, the solver processes the output currents at the end of each simulation (non-real-time), then it decides for the corrected voltage waveforms to be applied to the device that in turn will modify the output current of the next simulation (self-consistency). This procedure is iterated until the desired solution is achieved (iterative). HB allows to directly analyze the AC steady-state solution. On the other hand, no information is obtained about the transient response. HB does not introduce noise in the simulations due to the fact that the voltages applied to the device, obtained as solution of the network, are analytically generated as sine functions. On the other hand, the reduced number of frequency components included in the solution may limit the proper emulation of the voltage waveforms due to the network impedance. In the present code, the HB solver output does not discriminate between intrinsic and extrinsic currents/voltages. In such way, the HB solver can be used in combination with FDTD by simply using the extrinsic currents as input, and the extrinsic voltages as output to be fed to the FDTD/CMC code. An example of this hybrid mode is discussed in Chapter 8.

- **AC small-signal** (see Section 7.2): This is a non-self-consistent, post-processing, non-iterative solver. The network elements are included to the final output of the simulation in post-processing. Therefore, there is not interaction between the network and the intrinsic device behavior. (*i.e.* resistive elements would change the DC bias point of the intrinsic device). The network calculation and related extrinsic device performance are only based on an analytical model that assumes a system linear response.

  HB is self-consistently coupled with the device simulator code and it is specifically designed to take into account non-linear effects arising when the device is driven under large-signal operations.

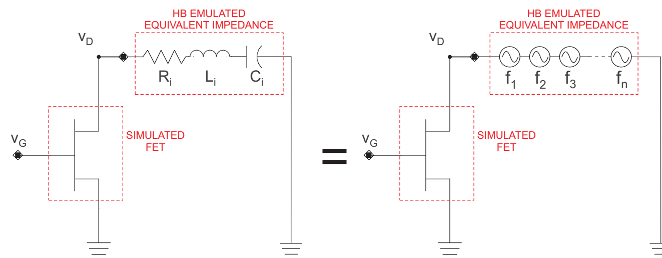

Figure 7.17: Series *RLC* resonant tank HB solver input mode.

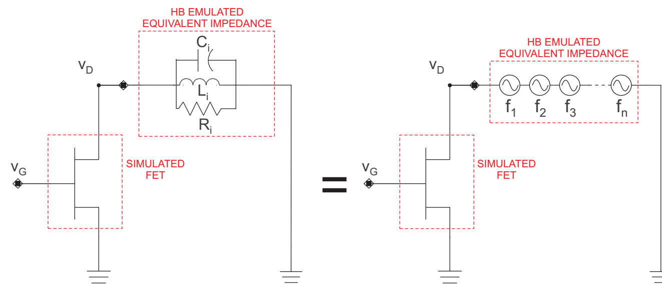

Figure 7.18: Parallel *RLC* resonant tank HB solver input mode.

HB Input Modes

The way HB is implemented in the present code is implicitly oriented toward the emulation of the equivalent network series impedance. Therefore, the different network models simply represent different "input modes" available to specify the equivalent network impedance at each specific i^{th} frequency component. The equivalent network series impedance, seen by each specific i^{th} frequency component, can be simply determined by the user through analytical calculations or by some intuitive network considerations (see Chapter 8.2 for an example). It is important to understand that the network topology associated to a certain input mode does not have to correspond to the actual network topology to be emulated.

The input mode associated to the topology shown in Figure 7.17 allows to specify the equivalent series impedance, seen by each specific i^{th} frequency component and to be solved by HB, as a series of a resistance (R), an inductor (L), and a capacitor (C) ($i.e.$ series resonant *RLC* circuit). This input mode is particular useful to emulate the

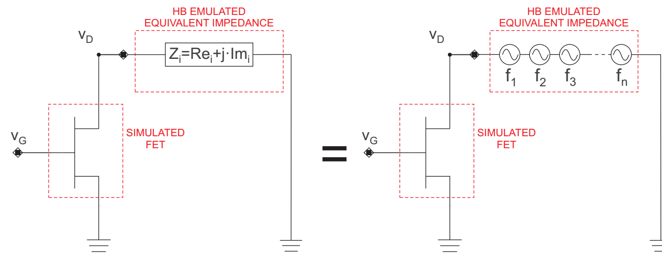

Figure 7.19: Manually-specified impedance HB solver input mode.

effect of circuit elements associated to a frequency-dependent impedance (*i.e.* reactive elements). The code will then automatically calculate the different complex impedance values associated to each specified frequency component as:

$$Z_i = R_i + j\left(\omega_i L_i - \frac{1}{\omega_i C_i}\right). \tag{7.37}$$

The input mode associated to the topology shown in in Figure 7.18 allows to specify the equivalent series impedance, seen by each specific i^{th} frequency component and to be solved by HB, as a shunt of a resistance (R), an inductor (L), and a capacitor (C) (*i.e.* parallel resonant RLC circuit connected in series to a device contact) as shown in Figure 7.18. This input mode is particular useful to emulate the effect of circuit elements associated to a frequency-dependent impedance (*i.e.* reactive elements). The code will then automatically calculate the different complex impedance value associated to each specified frequency. The input mode associated to the topology shown in in Figure 7.19 allows to manually specify the equivalent series impedance, seen by each specific i^{th} frequency component and to be solved by HB, in terms of its real and imaginary part as:

$$Z_i = Re\{Z_i\} + jIm\{Z_i\}. \tag{7.38}$$

This input mode is particular useful to emulate complex network topologies, where different impedance values are associated to each different frequency component, like, for instance, the output matching network of a Class-F Power Amplifier (see Section 2.5 and Section 8.2).

7.5 Conclusion

In this Chapter, we have reported the implementation of three different circuit solvers corresponding to the ones commonly available in circuit-level simulator tools. These three different solver options, combined with the Monte Carlo particle-based device simulation code, allow a complete and more realistic characterization of devices under DC, small-signal, and large-signal operations with a flexible approach enabling the user to manage the pros and cons related to these different circuit solver solutions.

Chapter 8

Large-signal Characterization of mm-wave GaN HEMT Power Amplifiers

8.1 Introduction

The drive towards device scaling and large output power levels in millimeter-wave power amplifiers [2] results in a highly non-linear, out-of-equilibrium charge transport regime of high-energy carriers. Particle-based Monte Carlo (MC) device simulators, in particular the ones based on the full band representation of the electronic structure and the phonon spectra, rather than analytical models, allow an accurate description of this carrier dynamics at the nanoscale. The small-signal AC analysis fails to analytically predict large-signal device performance, which must be assessed by simulating the device within the full range of the actual operating conditions [16]. In particular, when simulating a transistor used as power amplifier, the highly frequency selective high-Q matching network (required to suppress undesired harmonics generated due to the non-linearity typical of mm-band power amplifier classes such as -AB, -B, -F, and hard-driven Class-A described in Section 2.5) must be included in order to properly simulate the large voltage swing, due to the load-line [17], at the device output contacts as discussed in Section 2.5.

In this chapter, we exploit the self-consistently coupled Harmonic Balance / Monte Carlo simulator to perform circuit-device simulations by using an approach analogous to the one reported in literature for THz diode non-linear operations in mixers and frequency multiplier circuits [117, 118]. However, in our work, we perform for the first time a time-efficient accurate Monte Carlo large-signal numerical characterization of FETs used in RF power amplifier with high-Q matching network configuration by exploiting the benefits of the CMC, in terms of computational efficiency, and of the full-band approach, in terms accurate description of the high-energy carrier dynamics. The Monte Carlo techniques allows a more accurate simulation of the device behavior under large-signal operations than drift-diffusion simulators [119]. Furthermore, the

benefits of both the Harmonic Balance (HB) and the Finite Difference Time Domain (FDTD) circuit solvers are combined to include the effect of parasitic elements within a general approach that overcomes in such way the limitations and restrictive assumptions of the small-signal analysis. Such large-signal circuit-device simulations aim to provide for the first time a complete description of the critical non-linear effects in high-Q matched RF power amplifier [16, 17].

8.2 Time-efficient Simulation of mm-wave Power Amplifiers with High-Q Output Matching Networks: The Active Load-line Technique

A time domain FDTD solution (described in Section 7.3) of high-Q matching networks including large reactive elements requires a long simulation time due to the long transient response of the circuit (*i.e.* RC, L/C time constants, see Figure 7.10). However, this issue can be overcome by using an active-load line technique where a sinusoidal voltage generator tuned at the fundamental frequency is connected to the device output [13, 14, 16, 17]. This properly emulates the load-line drain voltage swing at the fundamental harmonic, and presents a short-circuit at the other harmonics, effectively emulating a high-Q matching network as shown in Figure 8.1. From the same figure, we can understand that not emulating the filtering effect of the output matching network results in a non-proper reproduction of the actual drain voltage stimulus of a FET used with such RF power amplifier topology reported in Section 2.5. For instance, in a Class-B amplifier, a half sinusoidal drain output current is associated to a full sinusoidal load-line voltage swing (*i.e.* drain voltage swing) due to the filtering effect of the matching network. Simulating the operation of a FET biased for Class-B operations by simply using a bias tee and a load connected to the device output (similarly to what shown in Figure 7.9) would result in a half sinusoidal drain voltage swing, which is not the case for a FET used in actual Class-B amplifiers. As previously discussed in Section 2.5 and high-lighted in Chapter 6, the drain voltage heavily affects the carrier distribution within the device.

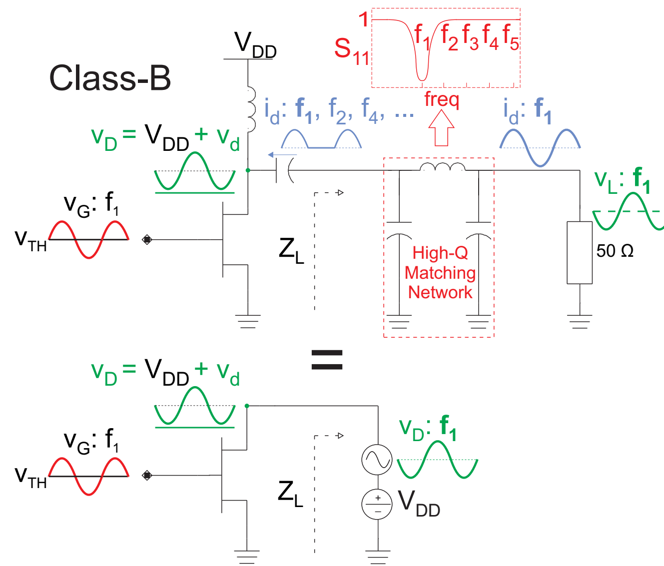

Figure 8.1: Example of a Class-B amplifier (*i.e.* conducting only for half of the input cycle) with high-Q matching network emulated by an active load-line technique.

The actual synthesized load impedance is determined in post-processing through Fourier analysis. The magnitude and phase of the complex load can be also adjusted, by adjusting the magnitude and the phase of the voltage generator, and the simulation and the subsequent impedance analysis can be iterated until the desired load impedance is obtained (see Figure 8.2). In such way, we can synthesize any chosen value of a complex impedance, and we can emulate a constant load for different input powers. This is needed because, when the FDTD solver is used, the drain voltage is given by the drain current swing and the load impedance value. On the other hand, when the active load-line technique is used, the value of the load impedance is given by the drain current swing and the drain voltage swing. So, in this latter case, the value of the emulated load impedance, for a certain fixed load-line drain voltage swing, changes as the drain current changes due to the increasing input power. Therefore, the drain voltage swing must be adjusted as the input power increases (*i.e.* increasing drain current) in order to preserve a constant amplitude and phase relation (*i.e.* the impedance complex value) between voltage and current at the device output contacts.

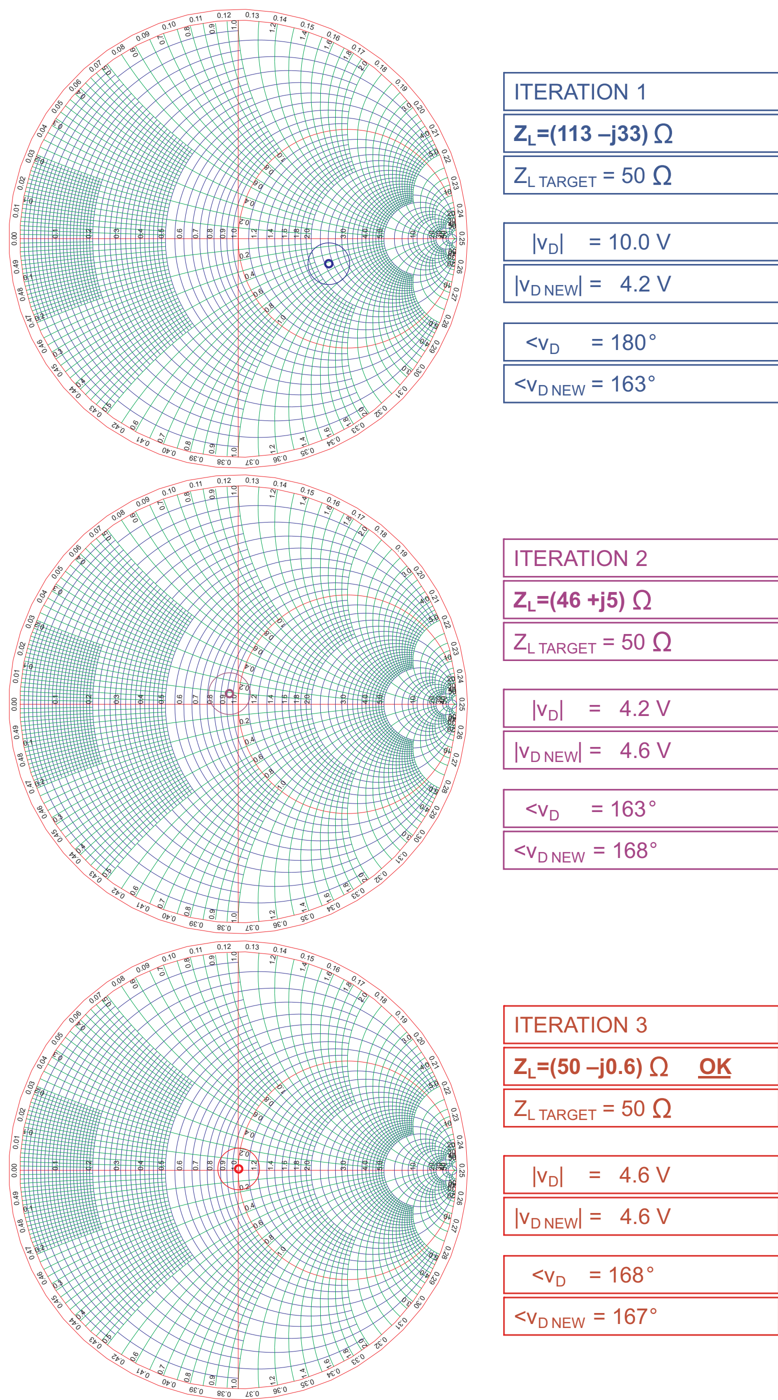

Figure 8.2: Smith chart plot of the load tuning iterative procedure for a device biased for Class-A operations at 25 GHz with emulated load impedance seen by the device at its output contacts of 50 Ω.

Furthermore, the ability to emulate and maintain a constant impedance avoids issues related to the fact that the active load-line itself feeds additional power to the device in addition to the gate input power. An example can clarify this issue: In a situation with a non-zero input gate voltage swing that results in a output drain current swing, a fixed drain voltage swing, generated by an active load-line, corresponds to a certain emulated load impedance seen by the device at its output contacts. Now, if we set a zero input gate voltage swing, we will still have a non-zero output drain current swing due to the combination of the active load-line drain voltage swing and the device output impedance (see Figure 2.11 for an example). In such case, we have a paradoxical situation of zero FET input power but non-zero FET output power simply because the active load-line is providing power to the system. However, if we apply the iterative procedure that adjusts the drain voltage swing to maintain a constant emulated load impedance, the active load-line drain voltage swing will be reduced at the next iteration due to the fact that the drain current output swing is reduced as result of the turn-off of the input gate voltage swing. This in turns will further reduce the drain current output swing, due to the now reduced active load-line drain voltage swing, which will eventually converge in a few iterations to a zero active load-line drain voltage swing, and zero drain current swing. Conversely, the active load line with a fixed voltage swing also improperly allows to emulate a negative resistance load impedance, which corresponds to a load-line drain voltage swinging from a low-current low-voltage point to a high-current high-voltage one in the IV space (see Section 2.5), without inducing any device oscillation. With such negative resistance load the FET would oscillate in a real situation, but with a constant load-line drain voltage swing this will not be the case. However, if we apply the iterative procedure that allows the drain voltage swing to be changed in order to maintain a constant emulated load impedance, an increase of such active load-line voltage swing leads to an increase of the output drain current, due to the device output impedance, properly resulting in an instable positive feedback

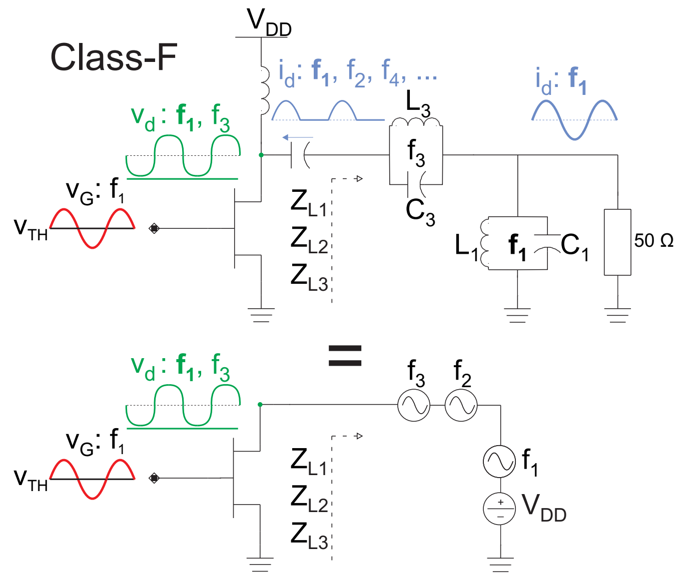

Figure 8.3: Example of a Class-F amplifier (*i.e.* conducting only for half of the input cycle and harmonic loading) with high-Q matching network and different load impedance values at the different harmonics emulated by active load-lines [13, 14].

and the consequent device oscillation. This will also occur with any non-negative load impedance located in the load instability region of the device.

It must be observed that, in case of simulations with a single-tone input, this active load-line iterative procedure simply represents a particular single-harmonic case of the general frequency-domain Harmonic Balance circuit solver described in Section 7.4. Thus, the HB circuit solver can be used to simulate the circuit topology connected to the device output, which corresponds to the bias tee and the load connected through the high-Q matching network, by simply solving a series impedance only specified at the fundamental frequency. Moreover, Class-F amplifiers can be also readily simulated with such method by simply solving different series impedance values only specified at the odd harmonics. This can be done by connecting different sinusoidal voltage generators tuned at the different odd harmonics as shown in Figure 8.3. Not connecting any voltage sinusoidal voltage generator tuned at the even harmonics corresponds to present a short-circuit at the even harmonics, effectively emulating the set-up of Class-F amplifiers (see Section 2.5). However, in the simulation of actual Class-F

amplifiers, the optimal impedance for odd harmonics will be generally large, while the optimal impedance for the even harmonics will be generally small and often including an imaginary part to cancel out the device output intrinsic capacitance [13, 14]. It must be noted that the convergence must be *simultaneously* achieved for each ith frequency component. Thus, the sinusoidal voltage generators must be continuously adjusted at *every* i^{th} frequency component as long as this condition is not met [13]. This because, in general, there is interaction between the frequency components in a non-linear system as discussed in Section 2.5 and 2.5.

The effect of the bias tee is emulated by simply setting the DC component at the drain voltage equal to the bias value. On the other hand, during the large-signal operations, the drain current DC component can be different than the DC bias value if for any reason the dynamic load-line is not symmetrical due to distortion or amplifier class of operations (*i.e.* Class-B self-biasing effect, see Section 2.5).

Unlike previous Monte Carlo codes coupled with time-domain circuit solvers used for FET power analysis [120, 121], our HB/CMC simulator allows a time-efficient simulation of devices connected the to high-Q matching networks required for the proper FET operations (especially for Class-B and Class-F). Due to the iterative nature of the HB algorithm, this FET simulation approach is possible only due to the computational efficiency of our CMC code as shown in from Figure 8.4. With our approach, the performance under large-signal operation can be evaluated through a physical device simulator, when the device stimuli appropriately mimic the actual device operating bench. All the large-signal out-of-equilibrium non-linear effects are intrinsically included in the nanoscale description of the carrier dynamics and accurately simulated by the CMC algorithm, which offers a more exact and physics based description than conventional drift and diffusion and hydrodynamic simulators.

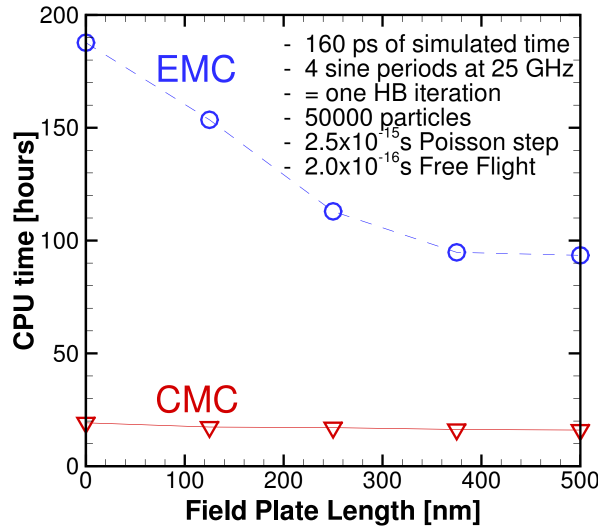

Figure 8.4: EMC and CMC simulation time (1HB iteration) vs increasing field plate length (i.e. decreasing peak electric field, decreasing peak scattering, decreasing scattering events, and decreasing number of CPU operations to be performed). The HB average number of iterations is 4. The simulations were run on a 64-bit Intel Xeon 3 GHz.

The High-Q Input Matching Network

As discussed in the previous paragraph, the output matching network, the load impedance, and the load-line voltage swing are key factors for the power amplifier operations, and they must be included in the simulation framework due to the fact that the large drain voltage swing strongly affects the carrier dynamics and the device operations in highly non-linear regime. Simulating a FET used as power amplifier without taking into account this aspect, by just applying a large input voltage swing without any load-line drain voltage swing (*i.e* a short-circuit load) represents a serious misconception.

The input matching network and the input signal generator impedance ("source impedance") can also play a quite important role and affect the device operations. The input voltage generator ("source generator") provides RF power to the FET. However, the input matching network and the source impedance change the actual input voltage

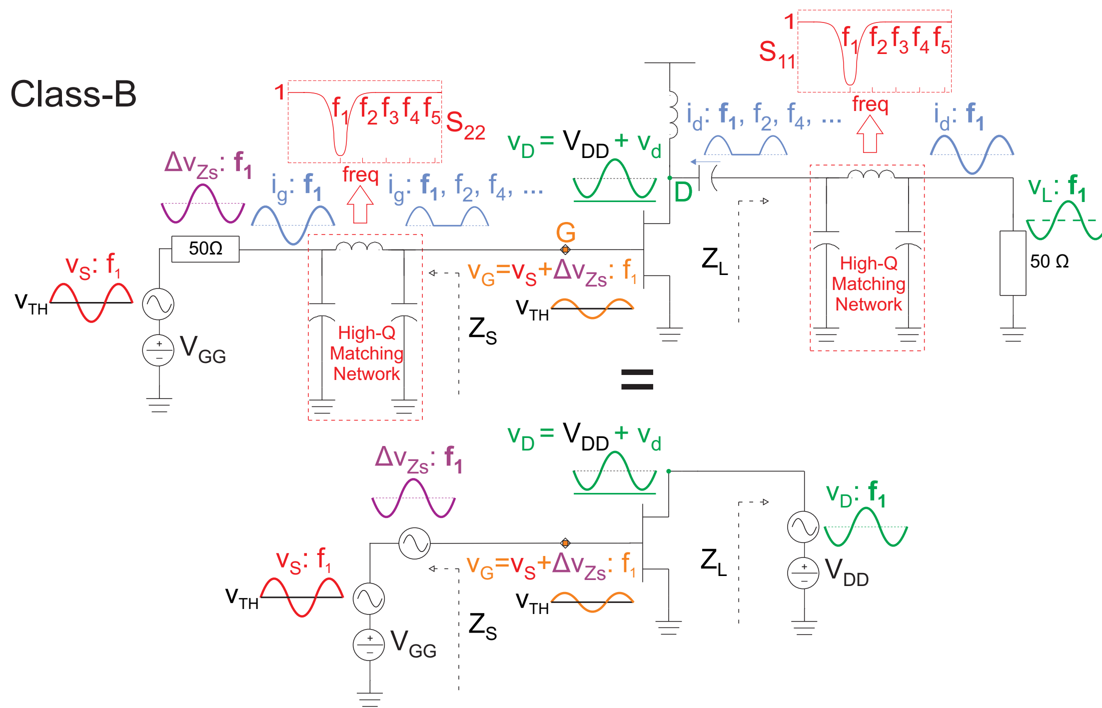

Figure 8.5: Example of a Class-B amplifier (*i.e.* conducting only for half of the input cycle) with input and output high-Q matching networks both of them emulated by an active load-line technique.

swing seen by the device at the gate contact. Ignoring the role of the output load-line corresponds to completely remove any voltage swing at the drain side. On the other hand, ignoring the role of the input matching network simply corresponds to provide the FET with a gate voltage swing that does not exactly correspond to the one that there would be in a real FET amplifier. Therefore, the impact of the input matching network and source impedance, in terms of reproduction of the actual device stimulus, is somehow less dramatic than the output matching network and the load impedance. In any case, the high-Q input matching network and source impedance can be easily included in our circuit-device simulation by simply realizing that they actually correspond to an ”input load-line”. From this viewpoint of view, the load is the source impedance, and the gate voltage and gate current at the fundamental frequency allow to obtain the actual source impedance seen by the device at its input contacts. Therefore, we can use the active load-line technique in this case as well. However, the fact that

a voltage generator (*i.e.* the RF input signal generator) is already connected to the device input might complicate this approach somehow. By virtue of the assumption of a passive and linear external network, we can however exploit the superposition principle and model the input load-line by simply adding a sinusoidal voltage generator tuned at the fundamental frequency. In this way, the actual gate voltage applied to the FET will be the sum of the input signal generator and the input load-line generator, which modifies the input signal, as shown in Figure 8.5. The emulated source impedance, Z_S, seen by the device at its input contacts is going to be the complex ratio in the frequency domain between the input load-line voltage generator and the fundamental frequency component of the gate current.

8.3 Inclusion of Parasitic Elements in Large-signal operations

The analytical inclusion of parasitic elements via post-processing based on small-signal analysis [53] (see Section 7.2) is not feasible in the large-signal regime. However, small value reactive parasitic elements can be efficiently simulated with the FDTD solver because the corresponding transient time response is much faster than the reactive elements associated with high-Q matching network elements (*i.e.* large reactive values). In such way, we can simulate the difference between the voltage applied to the device ”extrinsic” contacts and the voltage actually seen by the device at its ”intrinsic” contacts. The advantage of this real-time time-domain solution, unlike the small-signal AC analytical approach, is that there are not assumptions regarding the small-signal / large-signal regime and the linear / non-linear response of the device. One should note that these extrinsic parasitic elements can be simulated through the HB network solver algorithm as well, even though in such case higher harmonics beyond the fundamental one must be included in the solution as shown in Figure 7.14. For instance, as a rule of thumb, at least the first three harmonics should be included in order to account for the compression of the waveforms due to the power amplifier non-linear operations. As already said, the convergence must be *simultaneously* achieved for each ith frequency

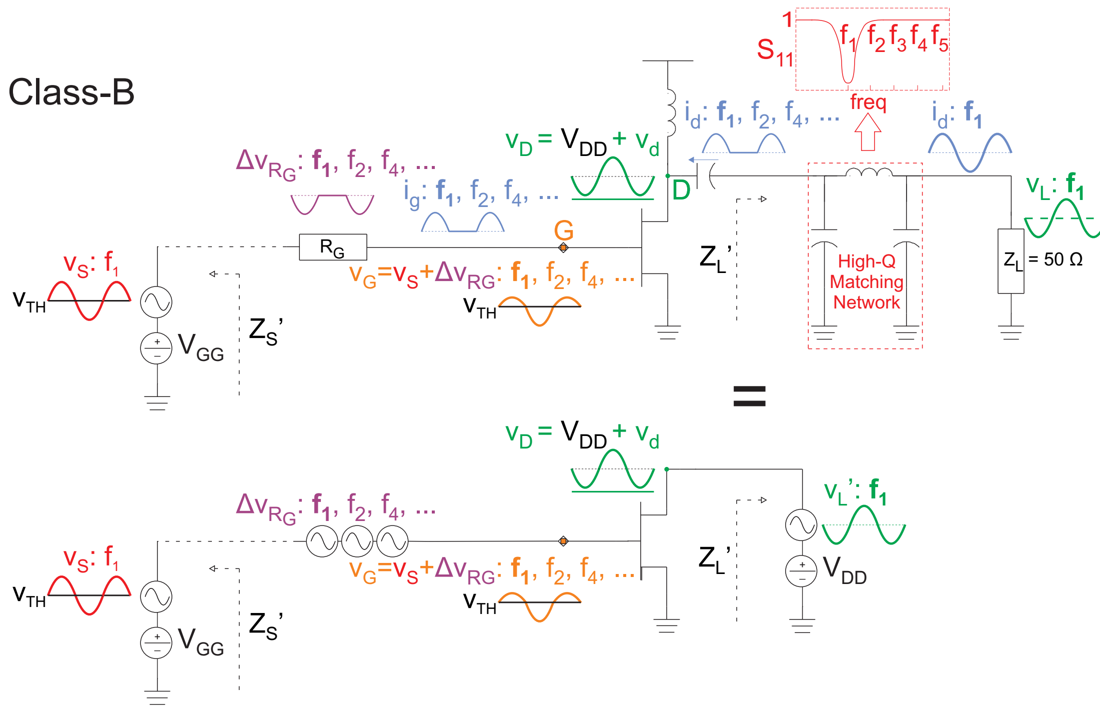

Figure 8.6: Example of a Class-B amplifier (*i.e.* conducting only for half of the input cycle) with high-Q matching network and parasitic gate resistance, R_G, both of them emulated by an active load-line technique.

component. Thus, the sinusoidal voltage generators must be continuously adjusted at *every* i^{th} frequency component as long as this condition is not met.

This HB approach of including parasitic elements might seem somehow problematic when the parasitic element is related to a contact where there is already a voltage generator connected, like the RF input generator at the gate or the active load-line at the drain. However, analogously to the approach used in the previous paragraph to emulate the effect of the input matching network, we can here use the superposition principle. For instance, we can simply emulate the voltage drop due to the parasitic gate resistance, R_G, by connecting an additional voltage generator to the gate contact. Then, the actual gate voltage applied to the gate will be superposition of the RF input signal and the voltage of this additional generator (*i.e.* minus the voltage drop across R_G). However, unlike the case of the input generator impedance, there is not a high-Q network between R_G and the device, so high harmonics beyond the fundamental must

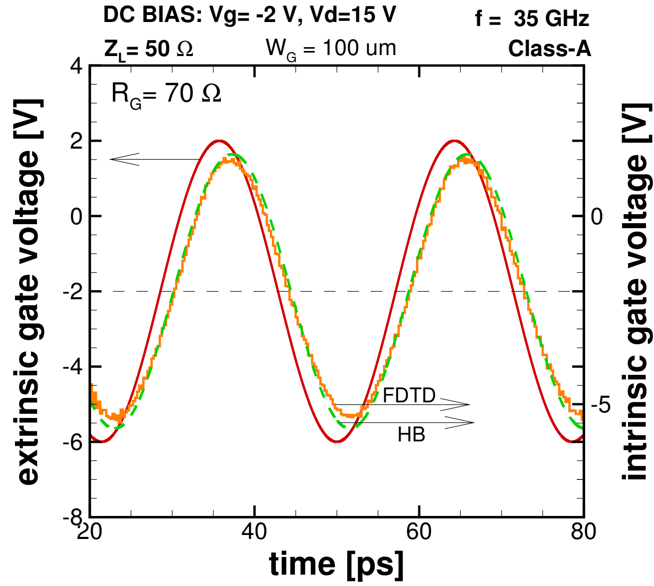

Figure 8.7: Effects of the gate resistance, R_G, on the intrinsic gate voltage waveform of a FET used as Class-A RF power amplifier at 35 GHz. R_G was simulated by using both the FDTD/CMC and HB/CMC approach.

be taken into account by adding more sinusoidal voltage generators tuned at these harmonic frequencies as shown in Figure 8.6. The HB algorithm will make sure that the complex ratio in the frequency domain (*i.e.* the impedance) at each single harmonic frequency between each of the voltage generators, each of them tuned at a specific frequency associated to one of the harmonics of the voltage drop across R_G as shown in Figure 8.6, and the gate current evaluated at each corresponding harmonic frequency will be equal to the R_G value. The equivalence between the inclusion of the effect of parasitic elements through FDTD or HB can be seen in Figure 8.7, where the intrinsic gate voltage waveform, resulting from the voltage drop across R_G, was calculated by using the two methods. Another example, related to a series resistance connected to the source contact, can be seen in Figure 8.19 in Section 8.5.

In such way, two different options are available to perform a complete large-signal characterization of an intrinsic device coupled with an external network of high-Q matching networks, source/load impedances, and parasitic elements: A fully HB network solver option and a hybrid HB/FDTD network solver option. In the latter case,

the HB algorithm emulates the high-Q matching network, the bias tee, and the load-line voltage swing, while the FDTD algorithm solves the extrinsic parasitic network. Both the options are self-consistently coupled to the CMC. On the other hand, the frequency domain network solution provided by HB allows the inclusion of the effects of those passive structures that are more easily characterized in the frequency domain than in the time domain. For instance, the HB algorithm can include the effect of a transmission line by using the S-parameters characterization, which still holds true even in large-signal regime for a passive and bias independent structure.

8.4 Large-signal Simulations of mm-wave GaN HEMT Power Amplifiers

The active load-line technique and the FDTD solver were combined with the CMC in order to include in the device large-signal operations the output high-Q matching network and the load impedance seen by the device at its output contacts (through the active load-line), and some relevant parasitic elements (through the FDTD solver). The simulated device is the one described in Chapter 6. The concept of dynamic load-line is described in Section 2.5. In the following examples, we used a constant active load-line voltage swing in order to characterize the different cases with the same drain voltage swing regardless of the different output current associated to different devices with different parasitic elements. This situation corresponds to different emulated load impedance values for the different device configurations. However, this approach allows us to perform a systematic comparison of different devices due to the fact that, as already discussed, the voltage value on the drain contact affects the device intrinsic behavior. Thus we set a constant active load-line swing amplitude in order reproduce the maximum drain voltage swing that keeps the device in the saturation region. This value was chosen by evaluating the static $I_D - V_D$ characteristic of a reference case where no parasitic elements nor reliability issues were included in the model. A different comparison method with a constant impedance approach (*i.e.* variable load-line voltage swing) is reported in the next section.

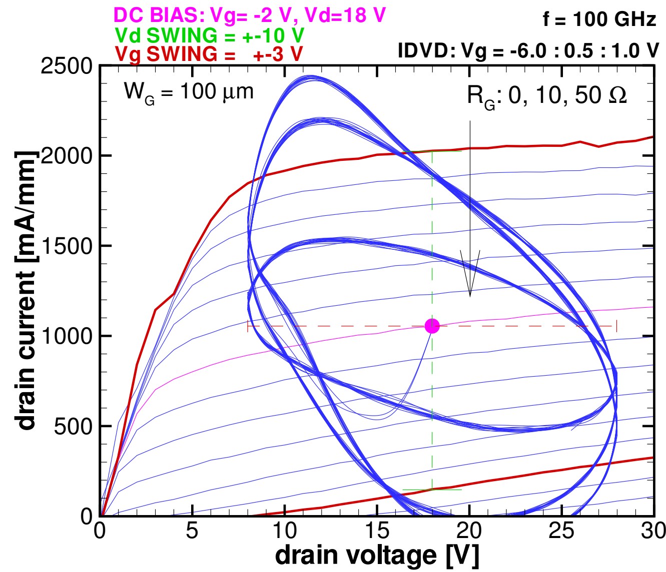

Figure 8.8: Effect of the gate parasitic resistance, R_G, on the dynamic load-line at 100 GHz.

It must be noted that the the waveforms and the dynamic load-lines reported in this and the next sections, are a smoothed version of the noisy output of the CMC simulator. However, this non-real-time smoothing was performed through post-processing only for graphic rendering purposes and does not introduce any phase delay due to the filtering. On the other hand, every data extraction, circuit solver calculation, and power performance evaluation were done processing the actual noisy simulation raw output.

The effect of the parasitic gate resistance (R_G) on very high-frequency (100 GHz) large-signal operations of a device biased in Class-A is shown in Figure 8.8. One can see that the voltage drop associated to the gate resistance leads to a lower gate voltage swing, seen by the device at its intrinsic gate contact, resulting in a lower output drain current swing. There is no change in the DC quiescent point of the device because the FET gate displacement current has only an AC component (no gate leakage due to tunneling was modeled). In this case, as already discussed, we chose to preserve a constant drain voltage swing regardless of the output drain current. This corresponds to a situation where devices with different R_G see a different load-impedance (*i.e.* the dif-

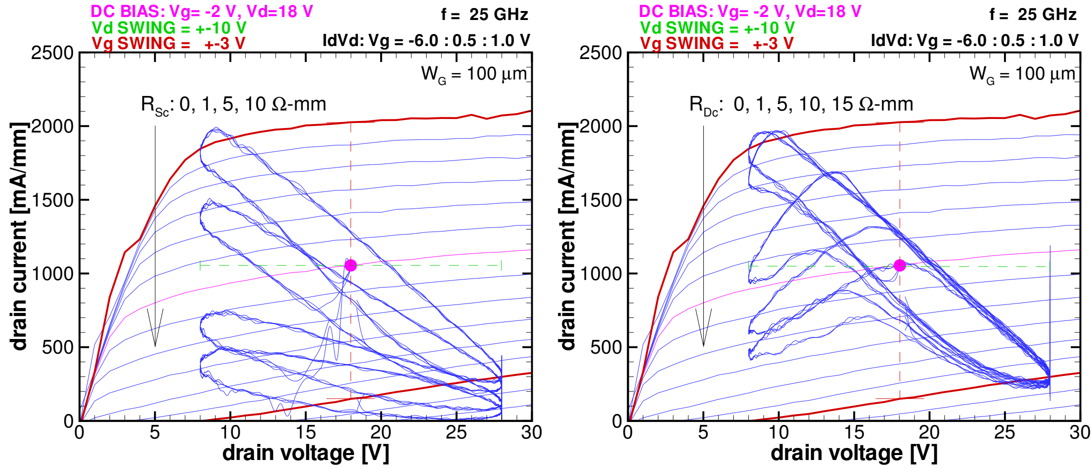

Figure 8.9: Effect of the source contact resistance R_{Sc} (left) and the drain contact resistance R_{Dc} (right) at 25 GHz with the device biased for Class-A operations. The overlaid DC $I_D - V_D$ corresponds to the $R_{Sc} = 0\ \Omega$ and $R_{Dc} = 0\ \Omega$ case.

ferent slope of the dynamic load-line major axis corresponds to the different real-part of the load impedance seen by the device). On the other hand, this approach allows us to evaluate the operation of different devices within the same range of drain voltage swing.

Using a similar approach, the impact of the parasitic source (R_{Sc}) and drain (R_{Dc}) contact resistance, in terms of reduced drain current (due to R_{Sc}) and higher drain saturation voltage, V_{Dsat}, (due to R_{Ds}), is reported, respectively, in the left and right side of Figure 8.9. As we can see, the effect of R_{Sc} is to shift the quiescent point by reducing the intrinsic v_{GS} seen by the device. This also results in a lower output current drain swing (*i.e.* difference between the maximum and minimum current achieved by the dynamic load-line) due to the reduced device extrinsic transconductance. Both this phenomena can be understood by observing the effect of a series source resistance on the DC characteristics as shown in Figure 7.5. On the other hand, R_{Dc} reduces the intrinsic v_D seen by the device causing a shift of the V_{Dsat} toward higher values as shown in Figure 7.7. This results in a reduced drain voltage swing range where the device dynamic load-line is above the saturation voltage. As we can see, as we

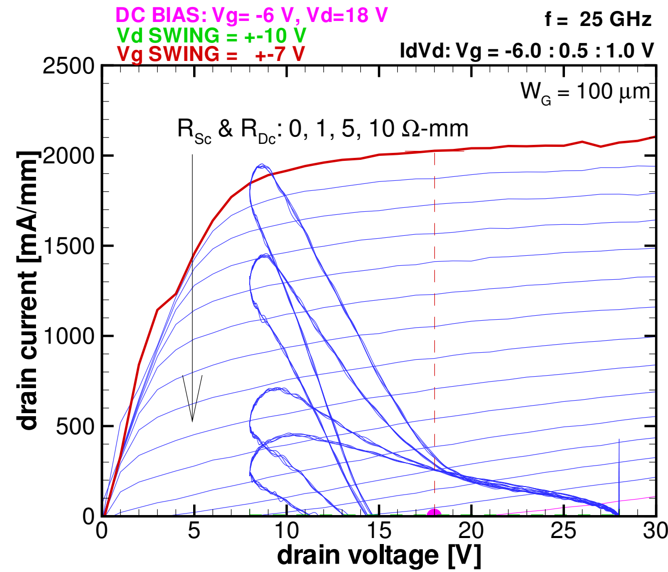

Figure 8.10: Effect of source and drain contact resistance at 25 GHz with the device biased for Class-B operations. The overlaid DC $I_D - V_D$ corresponds to the $R_{Sc} = 0\ \Omega$ and $R_{Dc} = 0\ \Omega$ case.

increase R_{Dc}, with a constant drain voltage swing, the dynamic load-line is distorted due to the fact that the device is not always operating within the saturation region. This reduces the output power as well as introduces the generation of spurious harmonics due to the large distortion. In such way, the single-tone input power is transferred from the fundamental harmonic to higher harmonics resulting in a reduced output power at the fundamental frequency. The combined effect of R_{Sc} and R_{Dc} on the load-line of a high-Q matched class-B amplifier can be seen in Figure 8.10. In general, the AC $I_D - V_D$ (see description later) sets the limits of the dynamic load-line large swing that are given by the drain saturation voltage (voltage lower bound), breakdown voltage (voltage upper bound), threshold voltage (current lower bound), and device maximum current (current upper bound). In must be noted that the "static" DC $I_D - V_D$ does not necessary correspond to the "dynamic" AC $I_D - V_D$ due to the fact that the device dynamically "moves" in the IV space during the AC operations . For instance, surface traps can lead to a difference between static and dynamic $I_D - V_D$ due to the DC-to-RF current dispersion described in Chapter 6.1.

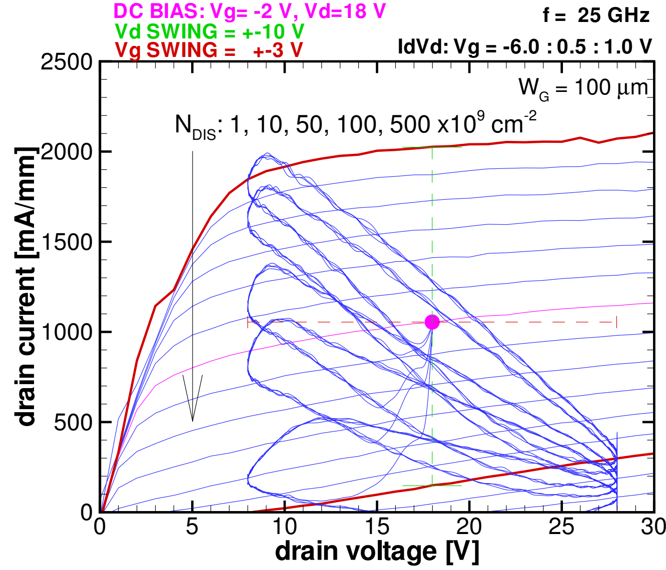

Figure 8.11: Effect of threading dislocation defects at 25 GHz with the device biased for Class-A operations. The overlaid DC $I_D - V_D$ corresponds to a dislocation density $N_{dis} = 0\ cm^{-3}$ case.

Reliability issues related to material defects, either due to fabrication issue or device stress degradation, such as threading dislocations and surface traps can be included by exploiting the accurate nanoscale description of the carrier dynamics description provided by the full band CMC approach. As we can see in the results shown in Figure 8.11, where increasing threading dislocation is modeled the simulation, the device behaves similarly to the R_{Sc} and R_{Dc} cases. This because the increased scattering randomizes the carrier velocity in the device access regions (*i.e.* reduced net velocity), resulting in larger source-gate and gate-drain access region series resistance. In this particular case, the effect of the source-gate access resistance is much larger than the gate-drain one. The effect of the latter can be only partially seen by the slight distortion introduced on the lowest dynamic load-line due to the increased drain saturation voltage that in turn reduces the extension of the saturation region available to the load-line swing.

Concerning the surface traps, they were modeled by using a sheet of negative charge depleting the 2-D electron gas (2DEG). This approach cannot take into account

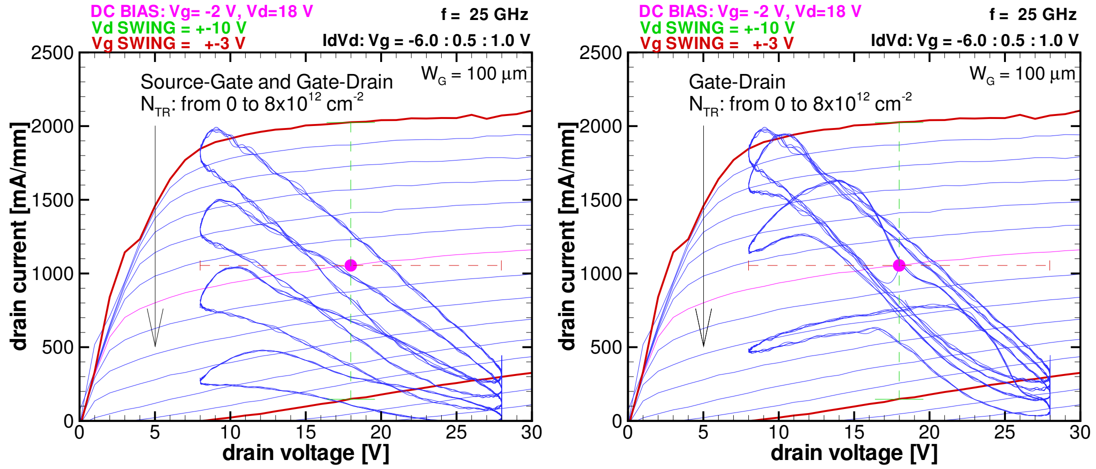

Figure 8.12: Effect of surface traps, located in both the source-gate and gate-drain regions (left) and only in the gate-drain region (right), at 25 GHz with the device biased for Class-A operations. The overlaid DC $I_D - V_D$ corresponds to a surface traps density $N_{TR} = 0\ cm^{-2}$ case.

the slow time-varying trapping/de-trapping process, but corresponds to the worst-case where all the traps are filled with electrons. As we can see from the left side of Figure 8.12, when the traps are located in both the source-gate and gate-drain regions, the device behaves similarly to the R_{Sc} and R_{Dc} cases and the threading dislocation case. The depletion of the 2DEG in the access regions increases the associated access region resistance. We can observe a slightly different behavior with respect to the threading dislocation case due to the underlying different physics of the two mechanisms. On the other hand, when the surface traps are located only in the gate-drain region, which corresponds to the high electric field region of the device where usually a large number of electrons are trapped during the device large-signal operations of GaN HEMTs biased with high drain voltages, we can observe a different behavior as shown in the right side of Figure 8.12. This behavior is somehow similar to the one reported for a series drain resistor shown in the right side of Figure 8.9. Again, we can explain this behavior as due to the increased access resistance in the gate-drain region due to the depletion of the 2DEG. In this case, the limitation of the drain voltage swing due to the increase drain saturation voltage is evident because there is not any increased

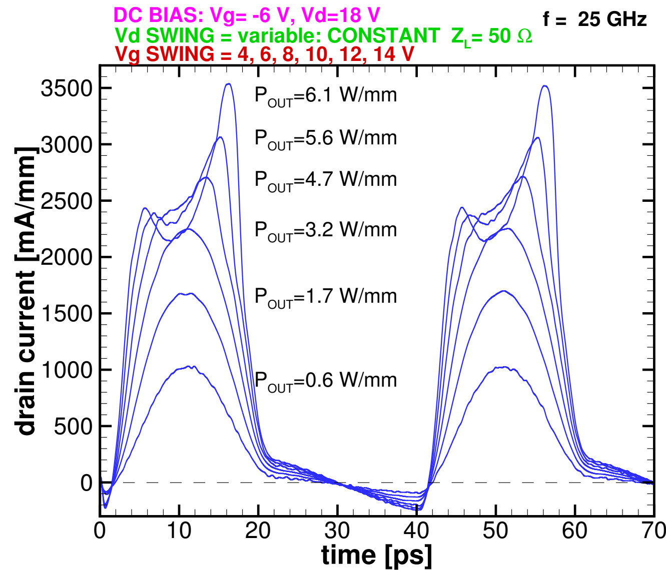

Figure 8.13: Output current waveforms for increasing input power at 25 *GHz* in Class-B operations with high-Q output matching network and emulated load $Z_L = 50\ \Omega$.

source-access region resistance masking this effect by reducing the transconductance. In fact, the dynamic load-lines exhibit pretty much the same major-axis slope (*i.e.* the small-signal transconductance is not affected), but the dynamic load-line are distorted when the drain voltage swing hits the lower limit set by the drain saturation voltage.

The distortion of Class-B output drain current waveforms, due to increasing input power, with constant load impedance of 50 Ω is shown in Figure 8.13. In this case, we used a constant load impedance / variable drain voltage swing approach. For each input power level, the CMC simulations were repeated each time adjusting the drain voltage swing, due to the load-line, according to the iterative procedure described in Section 8.1 and Section 7.4 applied to the fundamental frequency, until the desired load was achieved. In such way, we were able to perform a power sweep preserving a constant load impedance at the fundamental frequency (*i.e.* increasing drain voltage swing with increasing drain output current swing, and constant voltage-current phase relation).

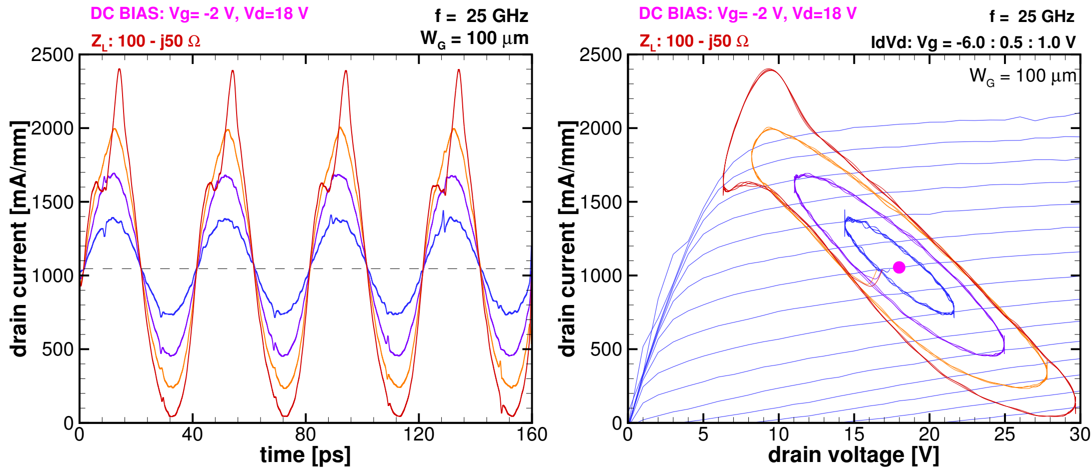

Figure 8.14: Output current waveforms (left) and related dynamic load-lines (right) for increasing input power with the device biased for Class-A with high-Q output matching network at 25 GHz.

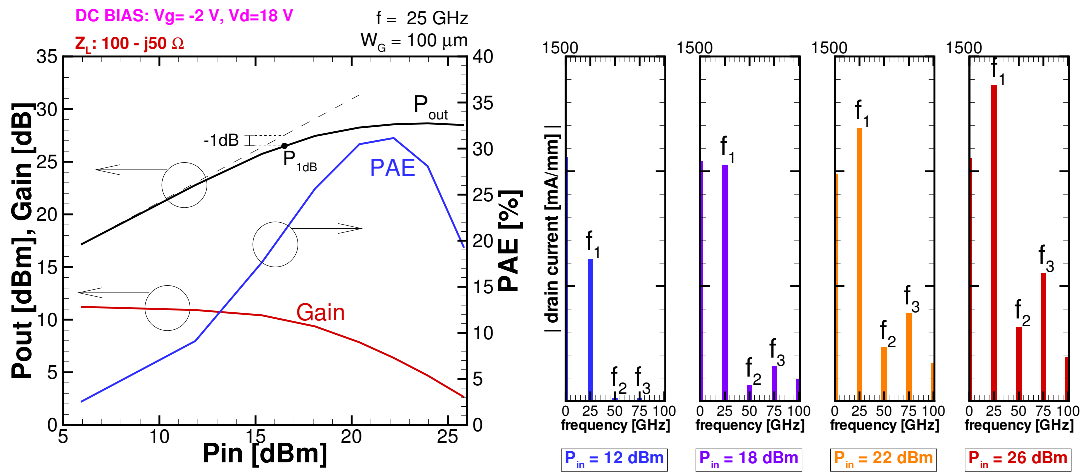

Figure 8.15: Output power (P_{out}), Power Added Efficiency (PAE), and Gain (left) and drain current harmonic content (right) versus input power (P_{in}) at 25 GHz with the device biased for Class-A with high-Q output matching network.

The output drain current waveforms of Class-A output drain current and the related dynamic load-lines for increasing input power with constant load impedance of $(100 - j50)\Omega$ are shown in Figure 8.14. In particular, we can also observe in this case, from the left side of Figure 8.14, that the increasing input power eventually introduces a distortion in the output waveforms. Moreover, we can verify that all the dynamic load-lines exhibit the same slope indicating the constant emulated load impedance.

Furthermore, as the input power increases, the ”expanding” shape of the dynamic load-lines can be related to the increasing output power (*i.e.* voltage-current swing product).

Finally, by exploiting this constant load-line power sweep, we can obtain the plot reported in the left side of Figure 8.15, that typically characterizes the performance of RF power amplifiers. As we can see, the increasing distortion corresponding to the increasing input power, transfers power from the fundamental to the higher harmonics and eventually reduces the power delivered to the load at the fundamental frequency. In particular, this results in the compression of the output waveforms and typically in the generation of a large third order harmonic component associated to this compressive behavior as clearly shown in the right histograms of Figure 8.15. The result is the saturation of the output power and the drop of the small-signal (*i.e.* low power) linear gain (G_{lin}), which is 11.2 dB in this case, as the input power increases. From this plot, we can extract typical figures of merit for RF power amplifiers as described in Section 2.5. We can see that this particular device achieves, with a load impedance of $100 - j50\ \Omega$ and a drain bias of 18 V, a peak output power of 33 dBm (corresponding to 7.3 W/mm) at a carrier frequency of 25 GHz. Moreover, the input referred 1dB compression point (P_{in1dB}) is 16.5 dBm, and output referred 1dB compression point (P_{out1dB}) is 26.5 dBm with a 1dB compression point gain (G_{1dB}) of 10.2 dB. The maximum power added efficiency (PAE_{max}) is 31% at $P_{in1dB} = 22\ dBm$.

The active load-line can be also used to investigate the large-signal output impedance of the device. This is accomplished by applying a constant bias voltage at the gate without applying any RF input. An example of a cause of non-linearity can be seen by observing the effect of the field plate length (L_{ARM}) on the large-signal output impedance (Z_{OUT}), due to the related drain-to-gate capacitance (C_{gd}), shown in Figure 8.16 in the Class-A bias configuratio. The C_{gd} values obtained from small-signal analysis are also reported in the inset. As we can see, the difference between the C_{gd} values at low (6 V) and high (18 V) drain voltage without field plate (*i.e.*

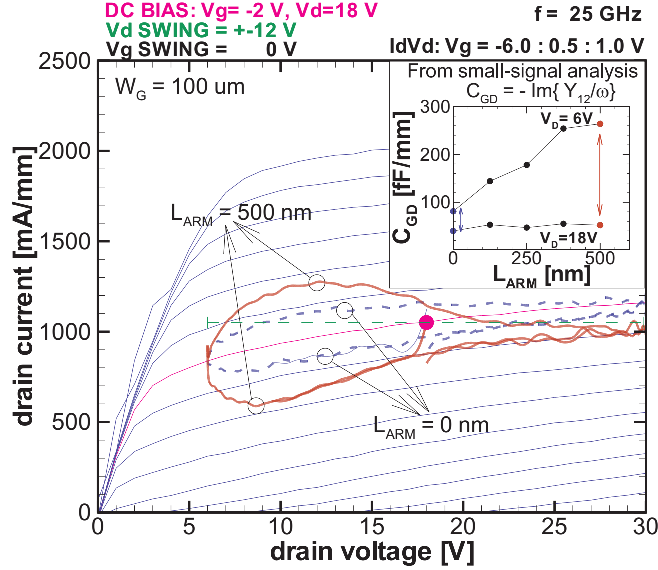

Figure 8.16: Impact of the field plate length (L_{ARM}) on the large-signal output impedance in a SiN passivated device at 25 GHz. The small-signal drain-to-gate feedback capacitance, C_{GD}, is also reported in the inset.

$L_{ARM} = 0$) is small ($\Delta C_{gd} = 41\ fF/mm$). On the other hand, there is a large difference ($\Delta C_{gd} = 211\ fF/mm$) between the low and high voltage values of C_{gd} in the full-length field plate case (*i.e.* $L_{ARM} = 500\ nm$). This because is due to the strong dependence of the gate-drain depletion region on the field plate length and the drain voltage (see Section 6.6). This results in a large variation of the output impedance, dominated by C_{gd} rather than C_{ds}, during the drain voltage swing with a full-length field plate as shown by the dynamic output impedance lines reported in the main plot. This large variation of C_{gd}, and the related large variation of the device intrinsic output impedance, results in large non-linearities during large-signal operations.

8.5 Characterization of a State-of-the-art mm-wave *InAlN/GaN* HEMT Power Amplifier

Exploiting the self-consistently coupled HB(FDTD)/CMC circuit-device simulator, we can simulate the large-signal performance of mm-wave FET power amplifiers and efficiently characterize the RF power performance of a state-of-the-art high-power, high-frequency *GaN* HEMT. In such way, we can relate parasitic elements, and material

defects to the large-signal power performance of mm-wave *GaN* HEMTs by simply including new/improved physical models in our CMC device simulator. Therefore, large-signal characterization can be readily and conveniently available during the device layout design phase. In particular, such kind of investigation can focus on the optimization of the intrinsic device layout in terms of real device issues such as parasitic elements (i.e. gate resistance, contact resistance, source/drain interconnection inductance), material defects (i.e. threading dislocations, surface traps, interface roughness), and electro-thermal effects. The goal of these large-signal Monte Carlo device simulations is to obviate the inability of properly evaluating analytically the large-signal performance, providing a TCAD tool for the device early-stage design flow. This is done by providing a computer-based performance evaluation in lieu of the extremely time-consuming and expensive iterations of prototyping and experimental large-signal characterization.

The typical characterization flow, if preliminary experimental DC and small-signal RF characterization are available for the device under investigation, starts with a fit of the experimental DC characteristics. Subsequently, the RF small-signal performance is assessed through AC/CMC simulations (by applying small step perturbations in order to extract the two-port network characterization of the device through Y-parameters as described in Chapter 2), and an agreement with the experimental measurements can be assessed. At this point, we finally proceed with the Power Amplifier predictive characterization through the HB(FDTD)/CMC code. Using this approach, we characterized the RF power performance of a state-of-the-art *InAlN*/*GaN* HEMT reported by Lee *et al.* [15, 16]. We first performed a fit of the experimental DC characteristics as show in the left side of Figure 8.17. Then, the RF small-signal performance were obtained through AC/CMC simulations and the agreement with the experimental results are shown in the right side of Figure 8.17. Despite a high f_T of 245 GHz, experimentally extracted after de-embedding the pad parasitic elements, this device has a

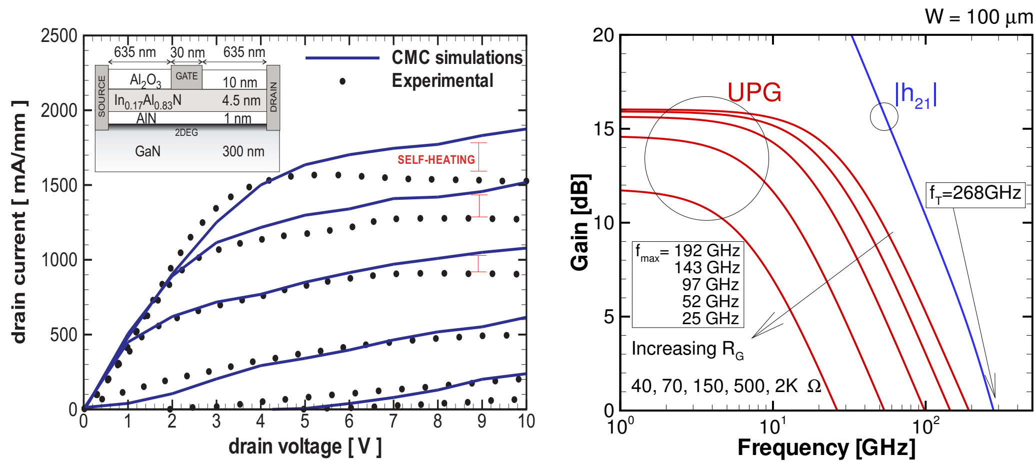

Figure 8.17: LEFT: Layout of the *InAlN/GaN* HEMT simulation domain and agreement between the DC simulated (without thermal correction) and experimental results [15]. RIGHT: AC small-signal Monte Carlo simulations. The experimental f_T is 245 GHz [15]. R_G was analytically embedded in the Y-parameters in post-processing by using the procedure described in Section 7.2.

relatively low f_{max} experimental value of 13 GHz due to the high gate resistance of the rectangular gate. On the other hand, mushroom-shaped gate structures or multifinger rectangular gates can significantly improve the f_{max} value as also discussed by Lee *et al.* in the same work. However, the effect of R_G cannot be directly taken into account by the solely Monte Carlo intrinsic device simulator. On the other hand, R_G, modeled as a lumped series resistor, can be included in our characterization by using the circuit-device simulation techniques described in Chapter 7 and in this chapter. In particular, the small-signal AC solver allows to embed R_G in the device small-signal characterization as shown in the right side of Figure 8.17, while the HB and the FDTD solvers, self-consistently coupled with the CMC, allow the inclusion of R_G under large-signal operations. Thus, in our simulated characterization we used different value of gate resistance, to assess the performance of this device with a low resistance gate structure.

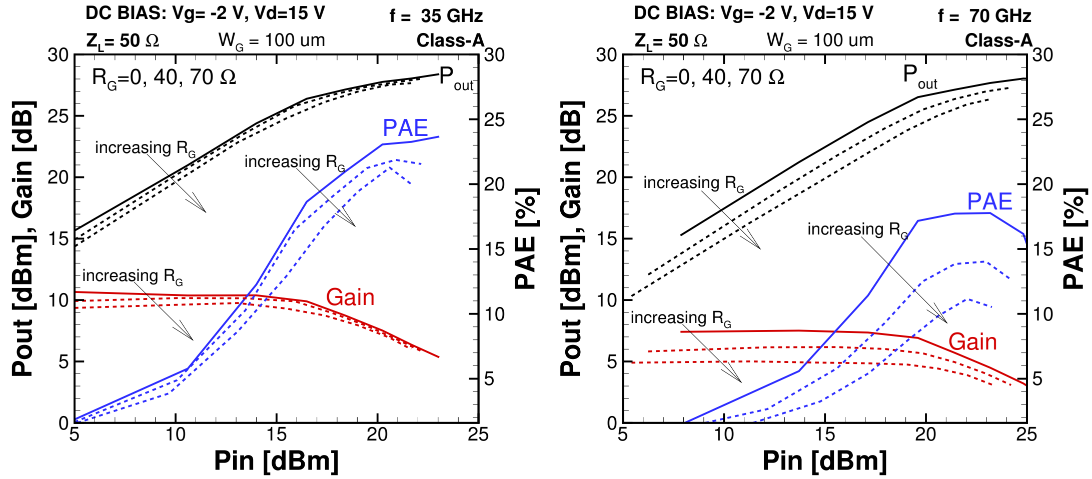

Figure 8.18: Effects of the gate resistance, R_G, on the 35 GHz (left) and 70 GHz (right) Class-A RF power amplifier performance of the device in Figure 8.17. R_G was simulated by using the FDTD/CMC approach.

The large-signal continuous-wave RF characterization was performed in the Ka-band at 35 GHz with the device biased for Class-A operations, and a 50 Ω load was emulated as example. The actual load that maximizes the output power can be obtained from experimental load-pull measurements as well as by computational load-pull technique [13] that can be readily obtained through the active load-line approach. Finally, we proceeded with the HB(FDTD)/CMC simulations including some real device issues such as parasitic elements and material defects.

The impact of the gate resistance, modeled as lumped elements, on the mm-wave power performance at 35 and 70 GHz is shown in Figure 8.18. As we can see in the $R_G = 0$ case, the higher is the frequency the lower is the device gain and resulting output power. Moreover, the effect of R_G is more pronounced at higher frequencies due to the larger displacement current flowing through the gate and the related larger voltage drop across R_G that in turn reduces the intrinsic gate voltage. Thus, even though R_G is not a reactive element, its impact is frequency dependent.

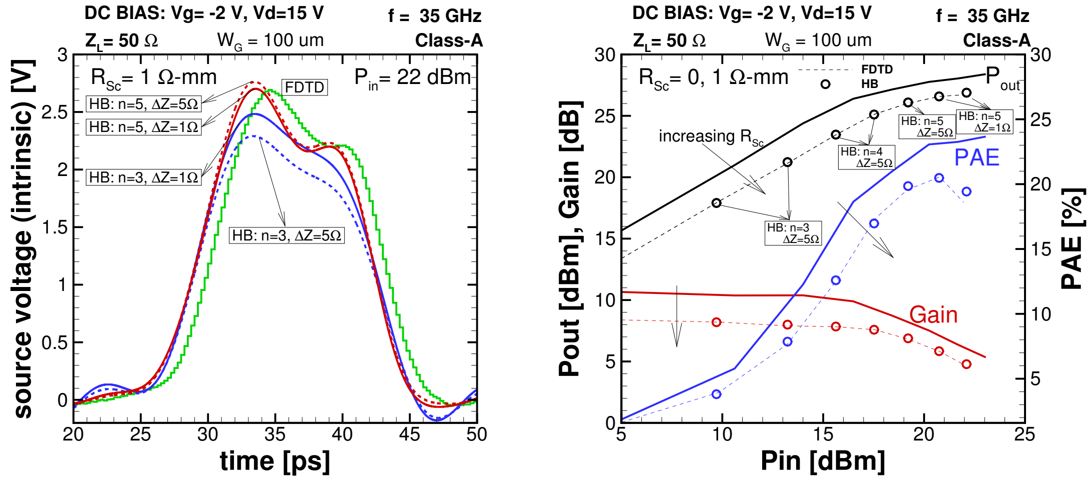

Figure 8.19: LEFT: Comparison between the FDTD and the HB circuit solvers in terms of intrinsic source voltage waveforms due to a series source contact resistor, R_{Sc}, for different number of harmonics, n, included in the solution (n + DC) and different maximum impedance error, ΔZ. The FDTD solver waveform delay is due to the real-time moving average filter. RIGHT: RF power amplifier performance characterization at 35 GHz for the Class-A operations obtained by simulating R_{Sc} through FDTD and HB, where different combinations of n and ΔZ were used, depending whether the convergence was achieved, for each different input power point.

In the left side Figure 8.19, we can verify the analogous results obtained by both the FDTD and HB circuit solvers when simulating a series source contact parasitic resistance, R_{Sc}. In particular, the accuracy of HB is given by the number of harmonics included in the simulation and the maxim error allowed between the desired and the emulated impedance value. In this example, a fairly good agreement between the two methods is achieved when at least 5 harmonics (in addition to the DC component), with a maximum impedance error of 1 Ω, are included. Moreover, we can verify that HB is able to properly reproduce the DC bias point change, due to the source contact series resistor, by simply using a sinusoidal voltage generator tuned at zero frequency. The power sweep reported in the right side Figure 8.19 confirms the equivalent power performance obtained through the two methods. On the other hand, including a large number of harmonics can be problematic for low input powers when the device operations are weakly non-linear. This because the HB solver is forced to evaluated an output

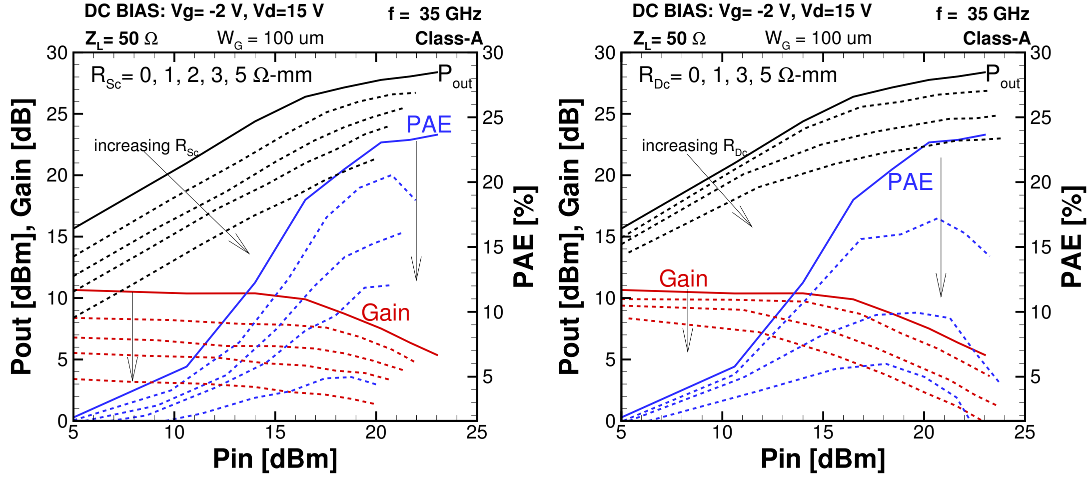

Figure 8.20: Effects of the source contact series resistance, R_{Sc}, (left) and the drain contact series resistance, R_{Dc}, (right) on the 35 GHz Class-A RF power amplifier performance of the device in Figure 8.17. R_{Sc} and R_{Dc} were simulated by using the FDTD/CMC approach.

current waveform with a very weak content of high harmonics (*i.e.* low signal-to-noise ratio), resulting in poor convergence of the HB algorithm. This issue is also shown in the right side Figure 8.19, where for low P_{in} only the simulations with a low number of harmonics, n, and a high impedance maximum error, ΔZ, achieved the convergence.

As we can see from Figure 8.20, both R_{Sc} and the series drain contact parasitic resistance, R_{Dc}, severely affect the power performance by reducing the amplifier gain. In particular, we notice that the largely reduced output power associated to large values of R_{Sc} increases P_{in1dB} (*i.e.* where the gain drops of 1 dB) due to the fact that a larger input power is required to have a dynamic load-line swing large enough to hit the $I_D - V_D$ limits given by the drain saturation voltage, threshold voltage, and device maximum current. However, this larger P_{in1dB} value is associated to a lower output power and gain. On the other hand, the effect of R_{Dc} is to heavily decrease P_{in1dB} through a higher drain saturation voltage (see Figure 8.9), which in turn seriously limits the load-line drain voltage swing and ultimately the output power by reducing the extension of the device linear operation region. The combined effects the gate and the source/drain parasitic resistances are reported in Figure 8.21.

Reference Case: P_{out} **= 26.7 dBm** at 35 Ghz with P_{in} = 17.3 dBm
P_{out} **= 24.6 dBm** at 70 Ghz with P_{in} = 17.3 dBm

P_{out} [dBm]	$R_{Sc}=R_{Dc}$=1 Ω-mm	$R_{Sc}=R_{Dc}$=2 Ω-mm	$R_{Sc}=R_{Dc}$=3 Ω-mm	$R_{Sc}=R_{Dc}$=5 Ω-mm
35 GHz, R_G=40Ω	23.9	21.7	20.1	17.4
35 GHz, R_G=70Ω	23.2	21.2	19.1	15.4
70 GHz, R_G=40Ω	19.3	17.4	15.9	12.9
70 GHz, R_G=70Ω	18	16.2	14.5	10.2

Figure 8.21: Output power (P_{out}) measured at for different combination of the gate, R_G, and the source/drain, R_{Sc} and R_{Dc}, parasitic resistances. P_{out} was measured at $P_{in} = 17.3\ dBm$, which is the input referred 1dB compression point in the reference case with no parasitic resistance that has $P_{out} = 26.7\ dBm$ at 35 GHz.

The impact of material defects such as threading dislocations and surface traps is shown Figure 8.22. As discussed before, the surface traps were modeled by using a static sheet of negative charge depleting the 2DEG. This approach cannot take into account the slow time-varying trapping/de-trapping process, but corresponds to the worst-case where all the traps are filled with electrons.

Concerning the effect of the dislocations, we can see on the left side of Figure 8.22 that their effect resembles the one of a series source and drain resistance due to the increased scattering along the channel. Increasing the dislocation density from $1 \cdot 10^9\ cm^{-3}$ to $500 \cdot 10^9\ cm^{-3}$ results in a 50% reduction of the output power and gain. This performance reduction is pretty much constant for increasing input power. On the other hand, effects of the surface traps, reported on the right side of Figure 8.22, are somehow different depending on their locations within the device. In particular, when the traps are located in both the source-gate and gate-drain regions, their effect is similar to the one of dislocations, which again can be described as an increased source-gate and gate-drain access region resistance due to the 2DEG depletion. However, when the negatively charged traps are only located in the gate-drain region, which is commonly the case due to the fact that most of the electrons are trapped in this region where the peak electric field is located, we can appreciate a performance reduction only for large input power. This can be explained with discussion similar to the ones reported in in

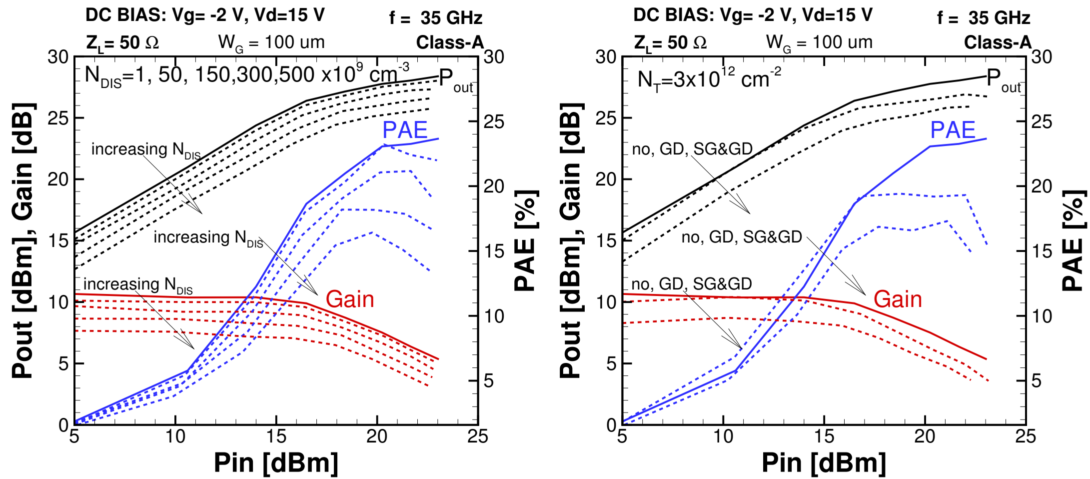

Figure 8.22: Effects of threading dislocations (left) and surface traps (right) on the 35 *GHz* Class-A RF power amplifier performance of the device in Figure 8.17. The traps were located in the gate-drain region (GD) and in both the source-gate and gate-drain regions (SG&GD).

Reference Case: P_{out} = 26.7 dBm at 35 Ghz with P_{in} = 17.3 dBm

P_{out} [dBm]	50×10^9	150×10^9	300×10^9	500×10^9
3×10^{12} (GD)	25.9	25.5	24.7	23.8
3×10^{12} (SG&GD)	25.7	24.9	24	23
6×10^{12} (GD)	25.1	24.5	23.7	22.7
6×10^{12} (SG&GD)	23.3	22.9	22.1	21.1

Columns: N_{DIS} [cm^{-3}]; rows: N_T [cm^{-2}]

Figure 8.23: Output power (P_{out}) measured at for different combination of dislocation density (N_{DIS}) and surface trap density (N_T). The traps were located in the gate-drain region (GD) and in both the source-gate and gate-drain regions (SG&GD). P_{out} was measured at $P_{in} = 17.3$ dBm, which is the input referred 1dB compression point in the reference case that has $P_{out} = 26.7$ dBm.

Section 8.4 concerning the dynamic load-line and a series drain resistor. The increased gate-drain access resistance increases the drain saturation voltage, limiting the maximum drain voltage swing due to the load-line. The combined effects of these material defects and reliability issues are reported in Figure 8.23.

Lastly, we compare the performance of the device when biased for Class-A and Class-B operations in Figure 8.24. As we can see, the Class-A performs better in terms of P_{out} due the fact that, as discussed in Section 2.5, Class-B amplifiers have an

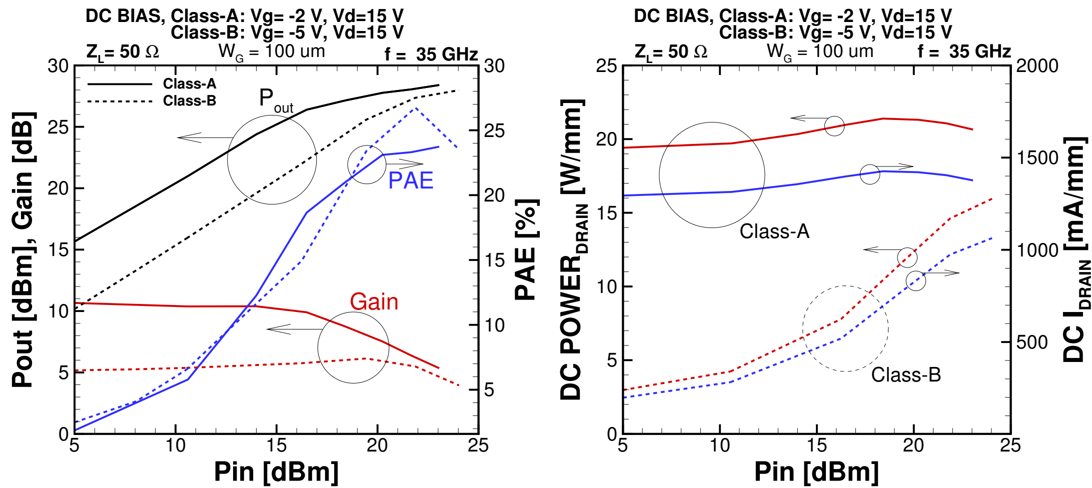

Figure 8.24: Comparison between the Class-A (solid) and Class-B (dashed) RF power amplifier performance of the device in Figure 8.17.

inherently lower gain than Class-A ones because a FET biased at the threshold voltage needs an input gate voltage swing which is larger that the one needed for a Class-A biasing to achieve the same peak current. On the other hand, despite the lower output power, the Class-B configuration achieves at $P_{in} = 21.9\ dBm$ better maximum PAE (about 27% versus 23%) due to the lower DC power consumption (about 14.7 W/mm versus 21 W/mm at such P_{in}) characteristic of Class-B amplifiers. Even in the low P_{in} range, the Class-B PAE is equal to the Class-A one despite the Class-B much lower output power. Moreover, we can see that the Class-B gain increases with increasing P_{in} from $G = 5.2\ dB$ to $G = 6.1\ dB$ (*i.e.* about +25%) due to self-biasing effect (see Section 2.5) typical of such configuration. In particular, we can see that while for Class-A the DC component of the drain current is pretty much constant with increasing P_{in}, the Class-B DC component of the drain current increases with P_{in}. The DC current corresponding to the maximum gain ($G = 6.1\ dB$ at $P_{in} = 19.3\ dBm$) is 800 mA/mm , which is equal to the one where the peak transconductance is achived in static $I_D - V_G$ characterization as experimentally reported [15].

8.6 Conclusion

In this chapter, the simulation of the large-signal performance of mm-wave FET power amplifiers with high-Q matching network was obtained for the first time through Full Band Monte Carlo particle-based device simulation self-consistently coupled with a Harmonic Balance frequency domain circuit solver. In particular, the combination of the active-load line technique, the fast CMC algorithm, the accuracy of the full band approach, and the parasitic element inclusion through a self-consistently coupled FDTD/CMC simulator allowed us to provide an accurate efficient numerical characterization of high-power, high-frequency GaN HEMTs under large-signal operations. The same set of simulations with time-domain and/or conventional Monte Carlo techniques would have required a prohibitive simulation time. On the other hand, this time-efficient HB/CMC circuit-device simulator also allows the rapid large-signal characterization of sub-millimeter band amplifiers, such as the ones based on InP devices with small-signal power gain cut-off frequency f_{max} as high as 1.2 THz [33], capable of efficiently operate at hundreds of gigahertz. For instance, the simulation of four time periods, corresponding to one HB iteration, of an InP-based amplifier operating at 250 GHz, assuming the same computational time of the devices described in this chapter, would require only two hours with the CMC technique on a 64-bit Intel Xeon 3 GHz.

Chapter 9

Conclusions

In the first chapters of this work, we provided a detailed study of the emerging technology of *N*-face *GaN* HEMTs, and showed the potentiality of this novel devices with respect to the traditional *Ga*-face transistors. Furthermore, short-channel effects were investigated in ultra-scaled *GaN* and *InP* HEMTs, and design guidelines based on nano-scale models were given to achieve superior performance in device operating in the millimeter and sub-millimeter band. Finally, the impact of the passivation dielectric and gate geometry on *GaN* HEMTs was studied, with particular emphasis on the link of the nano-scale carrier dynamics with the small-signal equivalent circuit parameters.

While the initial part of this work was mainly focused on DC and small-signal characterization, the last two chapters of this document are devoted to the development of simulation techniques that allow the large-signal characterization of *GaN* HEMTs used as power amplifiers in the millimeter-wave band. In particular, the aim of this effort was to allow the accurate simulations of the non-linear effects arising in FETs under large-signal operations, such as harmonic generation and gain compression, which cannot be predicted solely based on the DC and small-signal analysis. On the other hand, the critical effect of the matching network and the dynamic load-line on the power amplifier operations had to be included through circuit-device coupled simulations due to the lack of an analytical approach for such large-signal characterization. However, due to the fact that Monte Carlo particle-based simulations are traditionally extremely time-consuming, the challenge was to devise specific time-efficient techniques in order to obtain such circuit-device simulations with state-of-the-art devices in a reasonable time. This was accomplished through an active load-line technique and, with a more general approach, by self-consistently coupling for the first time a Harmonic Balance frequency domain circuit solver with a Full Band Cellular Monte Carlo. Our approach allowed an accurate characterization of high-Q matched *GaN* HEMT power amplifiers

through a particle-based device simulator in an efficient manner which has not been previously reported in the literature for other Monte Carlo codes.

The effect of parasitic elements was also included in order to account for an important aspect limiting the performance of high frequency real devices. This was accomplished with an approach that overcomes the limitation of the analytical parasitic inclusion of the small-signal analysis. In particular, two different options were implemented and self-consistently coupled to the Monte Carlo device simulator: A fully Harmonic Balance circuit solver and a hybrid Harmonic Balance / Time Domain Finite Difference circuit solver. The flexibility of this mixed approach turned out to be critical in order to overcome the limitation affecting both of these circuit solvers when coupled with an intrinsically noisy and relatively slow device simulation technique such as the particle-based Monte Carlo method. Our approach allowed for the first time a complete and realistic characterization of FET power amplifiers through a particle-based device simulator, including all the non-linearities resulting from large-signal operations. This was achieved by using a Monte Carlo code that is able to accurately describe the highly non-linear, out-of-equilibrium charge transport regime of high-energy carriers.

Future work will extensively use this code, which provided a TCAD tool for the device early-stage design, to perform computer-based performance evaluation in lieu of the extremely time-consuming and expensive iterations of prototyping and experimental large-signal characterization of millimeter and sub-millimeter wave devices.

[1] O. Ambacher, J. Majewski, C. Miskys, A. Link, M. Hermann, M. Eickhoff, M. Stutzmann, F. Bernardini, V. Fiorentini, V. Tilak, B. Schaff, and L. Eastman, "Pyroelectric properties of Al(In)GaN/GaN hetero and quantum well structures," *Journal of physics: condensed matter*, vol. 14, pp. 3399 – 3434, March 2002.

[2] U. Mishra, S. Likun, T. Kazior, and Y.-F. Wu, "GaN-Based RF power devices and amplifiers," *Proceedings of the IEEE*, vol. 96, no. 2, pp. 287–305, February 2008.

[3] A. Johnson, "Physical limitations on frequency and power parameters of transistors," *RCA Review*, vol. 26, pp. 163 – 177, 1965.

[4] Nidhi, S. Dasgupta, Y. Pei, B. Swenson, S. Keller, J. Speck, and U. Mishra, "N-polar GaN/AlN MIS-HEMT for Ka-Band power applications," *IEEE Electron Device Letters*, vol. 31, no. 12, pp. 1437 – 1439, December 2010.

[5] U. Mishra and J. Singh, *Semiconductor Device Physics and Design*, 1st ed. The Netherlands: Springer, 2008.

[6] L. Shen, S. Heikman, B. Moran, R. Coffie, N.-Q. Zhang, D. Buttari, I. Smorchkova, S. Keller, S. DenBaars, and U. Mishra, "AlGaN/AlN/GaN high-power microwave HEMT," *IEEE Electron Device Letters*, vol. 22, no. 10, pp. 457 – 459, October 2001.

[7] S. Gao, "High efficiency class-F RF/microwave power amplifiers," *IEEE Microwave Magazine*, vol. 7, no. 1, pp. 40 – 48, February 2006.

[8] F. Raab, "Class-f power amplifiers with maximally flat waveforms," *Microwave*

Theory and Techniques, IEEE Transactions on, vol. 45, no. 11, pp. 2007 – 2012, November 1997.

[9] O. Ambacher, J. Smart, J. R. Shealy, N. G. Weimann, M. M. K. Chu, W. J. Schaff, L. F. Eastman, R. Dimitrov, L. Wittmer, M. Stutzmann, W. Rieger, and J. Hilsenbeck, "Two-dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaN/GaN heterostructures," *Journal of Applied Physics*, pp. 3222–323, 1999.

[10] Nidhi, S. Dasgupta, Y. Pei, B. Swenson, D. Brown, S. Keller, J. Speck, and U. Mishra, "f_t and f_{max} of 47 and 81 GHz, respectively, on n-polar GaN/AlN MIS-HEMT," *IEEE Electron Device Letters*, vol. 30, no. 6, pp. 599 – 601, June 2009.

[11] F. Marino, N. Faralli, T. Palacios, D. Ferry, S. Goodnick, and M. Saraniti, "Effects of threading dislocations on AlGaN/GaN high-electron mobility transistors," *IEEE Transactions on Electron Devices*, vol. 57, no. 1, pp. 353–360, January 2010.

[12] T. Palacios, A. Chakraborty, S. Heikman, S. Keller, S. P. DenBaars, and U. K. Mishra, "AlGaN/GaN high electron mobility transistors with InGaN back-barriers," *IEEE Electron Device Letters*, vol. 27, no. 1, pp. 475–478, January 2006.

[13] O. Bengtsson, L. Vestling, and J. Olsson, "A computational load-pull method with harmonic loading for high-efficiency investigations," *Solid-State Electronics*, vol. 53, no. 1, pp. 86 – 94, 2009.

[14] A. Raffo, S. D. Falco, V. Vadala', and G. Vannini, "Characterization of GaN

HEMT low-frequency dispersion through a multiharmonic measurement system," *IEEE Transactions on Microwave Theory and Techniques*, vol. 58, no. 9, pp. 2490 – 2496, September 2010.

[15] D. S. Lee, J. Chung, H. Wang, X. Gao, S. Guo, P. Fay, and T. Palacios, "245-GHz InAlN/GaN HEMTs with oxygen plasma treatment," *IEEE Electron Device Letters*, vol. 32, no. 6, pp. 755–757, June 2011.

[16] D. Guerra, F. Marino, M. Saraniti, D. Ferry, S. Goodnick, and R. Soligo, "Large-signal mm-wave InAlN/GaN HEMT power amplifier characterization through self-consistent Harmonic Balance / Cellular Monte Carlo device simulation," *IEDM Technical Digest*, 2011.

[17] D. Guerra, F. Marino, D. Ferry, S. Goodnick, and M. Saraniti, "Large-signal full-band monte carlo device simulation of millimeter-wave power GaN HEMTs with the inclusion of parasitic and reliability issues." in *Proceedings of International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2011*, no. 87-90, September 2011.

[18] Y. Pei, C. Poblenz, A. Cordon, R. Chu, L. Shen, J. Speck, and U. Mishra, "X- and Ka-band power performance of AlGaN/GaN HEMTs grown by ammonia-MBE," *Electronics Letters*, vol. 44, no. 9, pp. 598–598, April 2008.

[19] R. Chu, L. Shen, N. Fichtenbaum, D. Brown, Z. Chen, S. Keller, S. DenBaars, and U. Mishra, "V-Gate GaN HEMTs for x-band power applications," *IEEE Electron Device Letters*, vol. 29, no. 9, pp. 974–976, September 2008.

[20] M. A. Khan, J. N. Kuznia, , D. T. Olson, W. J. Schaff, J. W. Burm, and M. S. Shur, "Microwave performance of a 0.25um gate AlGaN/GaN heterostructure

field effect transistor," *Applied Physics Letters*, vol. 65, no. 29, pp. 1121–1123, August 1994.

[21] J. Chung, W. Hoke, E. Chumbes, and T. Palacios, "AlGaN/GaN HEMT with 300-GHz $f_{\max}$," *IEEE Electron Device Letters*, vol. 31, no. 3, pp. 195–197, March 2010.

[22] Y.-F. Wu, M. Moore, A. Saxler, T. Wisleder, and P. Parikh, "40-W/mm double field-plated GaN HEMTs," in *Device Research Conference, 2006 64th*, June 2006, pp. 151–152.

[23] D. Guerra, M. Saraniti, N. Faralli, D. Ferry, S. Goodnick, and F. Marino, "Comparison of N-face and Ga-face GaN HEMTs through cellular monte carlo simulations," *IEEE Transactions on Electron Devices*, vol. 57, no. 12, pp. 3348–3354, December 2010.

[24] F. Marino, M. Saraniti, N. Faralli, D. Ferry, S. Goodnick, and D. Guerra, "Emerging N-face GaN HEMT technology: a cellular monte carlo study," *IEEE Transactions on Electron Devices*, vol. 57, no. 10, pp. 3327–3330, October 2010.

[25] D. Guerra, R. Akis, F. Marino, D. Ferry, S. Goodnick, and M. Saraniti, "Aspect ratio impact on rf and dc performance of state-of-the-art short-channel GaN and InGaAs hemts," *IEEE Electron Device Letters*, vol. 31, no. 11, pp. 1217–1219, November 2010.

[26] D. Guerra, M. Saraniti, D. Ferry, S. Goodnick, and F. Marino, "Carrier dynamics investigation on passivation dielectric constant and rf performance of millimeter-wave power GaN HEMTs," *IEEE Transactions on Electron Devices*, vol. 58, no. 11, pp. 3876–3884, November 2011.

[27] D. Guerra, F. Marino, S. Goodnick, D. Ferry, and M. Saraniti, “Extraction of gate capacitance of high-frequency and high-power GaN HEMTs by means of cellular monte carlo simulations,” *International Journal of High Speed Electronics and Systems*, 2011.

[28] D.-H. Kim and J. A. del Alamo, “Lateral and vertical scaling of $In_{0.7}Ga_{0.3}As$ HEMTs for post-Si-CMOS logic applications,” *IEEE Transactions on Electron Devices*, vol. 55, no. 10, pp. 2546 – 2553, October 2008.

[29] ——, “30-nm InAs pseudomorphic HEMTs on an InP substrate with a current-gain cutoff frequency of 628 GHz,” *Electron Device Letters, IEEE*, vol. 29, no. 8, pp. 830 – 833, August 2008.

[30] D. Guerra, F. Marino, R. Akis, D. Ferry, S. Goodnick, and M. Saraniti, “Terahertz-capability nanoscale InGaAs HEMT design guidelines by means of full-band monte carlo device simulations,” in *IEEE 11th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF), 2011*, January 2011, pp. 193 –196.

[31] D. Guerra, R. Akis, D. Ferry, S. Goodnick, M. Saraniti, and F. Marino, “Cellular monte carlo study of RF short-channel effects, effective gate length, and aspect ratio in GaN and InGaAs HEMTs,” in *14th International Workshop on Computational Electronics (IWCE), 2010*, October 2010, pp. 1 – 4.

[32] D.-H. Kim and J. A. del Alamo, “30-nm InAs PHEMTs with f_T = 644 GHz and f_{max} = 681 GHz,” *Electron Device Letters, IEEE*, vol. 31, no. 8, pp. 806 – 808, August 2010.

[33] R. Lai, X. B. Mei, W. R. Deal, W. Yoshida, Y. M. Kim, P. H. Liu, J. Lee, J. Uyeda,

V. Radisic, M. Lange, T. Gaier, L. Samoska, and A. Fung, “A sub 50 nm InP HEMT device with fmax greater than 1 THz,” *IEEE International Electron Devices Meeting*, December 2007.

[34] D. Guerra, F. Marino, R. Akis, S. Goodnick, D. Ferry, and M. Saraniti, “Carrier dynamics study on transit time and frequency response in highly scaled HEMTs through full-band monte carlo device simulations,” in *17th International Conference on Electron Dynamics in Semiconductors, Optoelectronics and Nanostructures (EDISON), 2011*, August 2011.

[35] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*. Wien, NewYork: Springer–Verlag, 1989.

[36] C. Hamaguchi, *Basic Semiconductor Physics*. Springer, Corrected edition, 2006.

[37] G. Meneghesso, F. Rampazzo, P. Kordos, G. Verzellesi, and E. Zanoni, “Current collapse and high-electric-field reliability of unpassivated GaN/AlGaN/GaN HEMTs,” *IEEE Transactions on Electron Devices*, vol. 53, pp. 2932–2941, 2006.

[38] M. Fischetti and S. Laux, “Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects,” *Physical Review B*, vol. 38, no. 14, pp. 9721–9745, Nov. 1988.

[39] M. Saraniti and S. Goodnick, “Hybrid full-band Cellular Automaton/Monte Carlo approach for fast simulation of charge transport in semiconductors,” *IEEE Transactions on Electron Devices*, vol. 47, no. 10, pp. 1909–1915, October 2000.

[40] N. W. Ashcroft and N. Mermin, *Solid State Physics*. Tokyo: Holt–Sauders International Editions, 1981.

[41] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1976, 5th ed., 1976.

[42] J. C. Phillips, "Energy-band interpolation scheme based on a pseudopotential," *Physical Review*, vol. 112, pp. 685–695, Nov. 1958.

[43] J. C. Phillips and L. Kleinman, "New method for calculating wave functions in crystals and molecules," *Physical Review*, vol. 116, pp. 287–294, Oct. 1959.

[44] S. Yamakawa, R. Akis, N. Faralli, M. Saraniti, and S. M. Goodnick, "Rigid ion model of high field transport in GaN," *Journal of Physics: Condensed Matter*, vol. 21, no. 17, pp. 174 206–174 222, 2009.

[45] S. Yamakawa, J. Branlard, M. Saraniti, and S. M. Goodnick, "High-field transport in GaN and AlGaN/GaN heterojunction field effect transistors," in 14th *International Conference on Nonequilibrium Carrier Dynamics in Semiconductors (HCIS-14)*, M. Saraniti and U. Ravaioli, Eds., no. 110. Springer Proceedings in Physics, 2006, pp. 133–138.

[46] W. Hackbusch, "The frequency decomposition multi-grid algorithm," in *Robust Multi–Grid Methods. Proceedings of the Fourth GAMM–Seminar, Kiel, January 22 to 24, 1988*, ser. Notes on numerical fluid mechanics, W. Hackbusch, Ed., vol. 23. Braunschweig / Wiesbaden: Fried. Vieweg & Sohn, 1989, pp. 96–105.

[47] D. Ferry, "The onset of quantization in ultra-submicron semiconductor devices," *Superlattices and Microstructures*, vol. 27, pp. 61–66, 2000.

[48] I. Knezevic, D. Vasileska, and D. Ferry, "Impact of strong quantum confinement on the performance of a highly asymmetric device structure: Monte Carlo particle-based simulation of a focused-ion-beam mosfet," *IEEE Transactions on Electron Devices*, vol. 49, pp. 1019–1026, 2002.

[49] D. M. Pozar, *Microwave Engineering*, 3rd ed. Wiley, 2005.

[50] R. Hockney and J. Eastwood, *Computer Simulation Using Particles*. Bristol: Adam Hilger, 1988.

[51] S. Laux, "Techniques for small-signal analysis of semiconductor devices," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 10, pp. 2028–2037, Oct 1985.

[52] R. A. Warriner, "Computer simulation of gallium arsenide field-effect transistors using Monte-Carlo methods," *IEE Solid State and Electron Devices*, vol. 1, no. 4, pp. 105–110, July 1977.

[53] S. Babiker, A. Asenov, N. Cameron, S. Beaumont, and J. Barker, "Complete Monte Carlo RF analysis of "Real" short-channel compound FET's," *IEEE Transaction on Electron Devices*, vol. 45, no. 8, pp. 1644–1652, Aug 1998.

[54] G. Gonzalez, *Microwave Transistor Amplifiers: Analysis and Design*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 1996.

[55] R. Akis, J. Ayubi-Moak, N. Faralli, D. K. Ferry, S. M. Goodnick, and M. Saraniti, "The upper limit of the cutoff frequency in ultrashort gate-length InGaAs/InAlAs HEMTs: A new definition of effective gate length," *IEEE Electron Device Letters*, vol. 29, no. 4, pp. 306–308, April 2008.

[56] K. N. S.M. Sze, *Physics of Semiconductor Devices*, 3rd ed. New Jersey: John Wiley & Sons, 2007.

[57] G. Crupi, D.-P. Schreurs, A. Raffo, A. Caddemi, and G. Vannini, "A new millimeter-wave small-signal modeling approach for pHEMTs accounting for the output conductance time delay," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 56, no. 4, pp. 741–746, April 2008.

[58] F. Diamant and M. Laviron, "Measurement of extrinsic series elements of a microwave MESFET under zero current conditions," in *Proceedings European Microwaves Conference,Helsinki, Finland*, September 1982, pp. 451–456.

[59] R. Anholt and S. Swirhun, "Equivalent-circuit parameter extraction for cold GaAs MESFETs," *IEEE Transactions on Microwave Theory and Techniques*, vol. 39, no. 7, pp. 1245–1247, July 1991.

[60] G. Crupi, D. Xiao, D. M. M.-P. Schreurs, E. Limiti, A. Caddemi, W. De Raedt, and M. Germain, "Accurate multibias equivalent-circuit extraction for GaN HEMTs," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 10, October 2006.

[61] S. A. Maas, *Nonlinear Microwave and RF Circuits, 2nd ed.* Norwood, MA: Artech House, 2003.

[62] R. Gilmore and L. Besser, *Practical RF Circuit Design for Wirelss Systems, Volume II*. Norwood, MA: Artech House, 2003.

[63] M. Bailey, "Intermodulation distortion in pseudomorphic HEMTs and an extension of the classical theory," *EEE Transactions on Microwave Theory and*

Techniques, vol. 48, no. 1, pp. 104 – 110, January 2000.

[64] E. Srinidhi, R. Ma, G. Kompa, and A. Bangert, "Device inherent IMD reproducibility issues in GaN HEMTs," in *European Microwave Integrated Circuits Conference, 2009. EuMIC 2009.*, September 2009, pp. 108 – 111.

[65] A. Goldsmith, *Wireless Communications*. New York, NY: Cambridge University Press, 2005.

[66] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed. John WILEY and Sons, 2007.

[67] M. Sumiya, K. Yoshimura, T. Ito, K. Ohtsuka, S. Fuke, K. Mizuno, M. Yoshimoto, H. Koinuma, A. Ohtomo, and M. Kawasaki, "Growth mode and surface morphology of a GaN film deposited along the N-face polar direction on c-plane sapphire substrate," *Journal of Applied Physics*, vol. 88, pp. 1158–1165, 2000.

[68] S. Rajan, A. Chini, M.H.Wong, J. Speck, and U. Mishra, "N-polar GaN/AlGaN/GaN high elecron moobility transistors," *Journal of Applied Physics*, vol. 102, p. 044501, 2007.

[69] S. Keller, C. Suh, Z. Chen, R. Chu, N. Fichtenbaum, M.Furukawa, S. DenBaars, J. Speck, and U. Mishra, "Properties of N-polar AlGaN/GaN heterostructures and field effect transistors grown by metalorganic chemical vapour deposition," *Journal of Applied Physics*, vol. 103, p. 033708, 2008.

[70] J. Chung, E. Piner, and T. Palacios, "N-face GaN/AlGaN HEMTs fabricated through layer transfer technology," *IEEE Electron Device Letters*, vol. 30, no. 2, pp. 113–116, February 2009.

[71] M. H. Wong, Y. Pei, T. Palacios, L. Shen, A. Chakraborty, L. S. McCarthy, S. Keller, S. P. DenBaars, J. S. Speck, and U. K. Mishra, "Low nonalloyed ohmic contact resistance to nitride high electron mobility transistors using N-face growth," *Applied Physics Letters*, vol. 91, pp. 232–233, 2007.

[72] K. Xu and A. Yoshikawa, "Effects of film polarities on InN growth by molecular-beam epitaxy," *Applied Physics Letters*, vol. 83, pp. 251–253, 2003.

[73] M. Wong, Y. Pei, R.Chu, S. Rajan, B. Swenson, D. Brown, S. Keller, S. DenBaars, J. Speck, and U. Mishra, "N-face metal-insulator-semiconductor high-electron-mobility transistors with AlN back-barrier," *IEEE Electron Device Letters*, vol. 29, no. 10, pp. 1101 – 1104, October 2008.

[74] S. Rajan, C. S. Suh, J. S. Speck, and U. K. Mishra, "N-polar aluminum gallium nitride nitride/gallium nitride enhancement-mode field effect transistor," US Patent PCT/US2006/036 377, 2007.

[75] U. K. Mishra, "Status of AlGaN/GaN HEMT technology - A UCSB perspective," 13th *GAAS Symposium*, 2005.

[76] U. Singisetti, M. H. Wong, S. Dasgupta, Nidhi, B. Swenson, B. Thibeault, J. Speck, and U. Mishra, "Enhancement-mode N-polar GaN MISFETs with self-aligned source/drain regrowth," *IEEE Electron Device Letters*, vol. 32, no. 2, pp. 137 – 139, February 2011.

[77] S. Kolluri, Y. Pei, S. Keller, S. Denbaars, and U. Mishra, "RF performance of N-polar AlGaN/GaN MIS-HEMTs grown by MOCVD on sapphire substrate," *IEEE Electron Device Letters*, vol. 30, no. 6, pp. 584–586, June 2009.

[78] C. K. Birdsall and D. Fuss, “Clouds-in-clouds, clouds-in-cells physics for many-body plasma simulation,” *Journal of Computational Physics*, vol. 135, no. 2, pp. 141–148, August 1997.

[79] S. Pearton, F. Ren, A. Zhang, , and K. Lee, “Fabrication and performance of GaN electronic devices,” *Materials Science and Engineering*, vol. 30, pp. 55 – 212, 2000.

[80] K. Kim, W. Lambrecht, and B. Segall, “Elastic constants and related properties of tetrahedrally bonded BN, AlN, GaN, and InN,” *Physical Review B*, vol. 56, p. 7018, 1997.

[81] O. Ambacher, B. Foutz, J. Smart, J. Shealy, N. Weimann, K. Chu, M. Murphy, A. Sierakowski, W. Schaff, L. Eastman, R. Dimitrov, A. Mitchell, and M. Stutzmann, “Two dimensional electron gases induced by spontaneous and piezoelectric polarization in undoped and doped AlGaN/GaN heterostructures,” *Journal of Applied Physics*, vol. 87, pp. 334–344, 2000.

[82] U. Karrer, O. Ambacher, , and M. Stutzmann, “Influence of crystal polarity on the properties of Pt/GaN schottky diodes,” *Applied Physics Letters*, vol. 77, pp. 2012–2014, 2000.

[83] F. Schwierz and J. J. Liou, *Modern Microwave Transistors, Theory, Design and Performance*. Wiley InterScience, 2003.

[84] S. Rajan, M. Wong, Y. Fu, F. Wu, J. S. Speck, and U. K. Mishra, “Growth and electrical characterization of N-face AlGaN/GaN heterostructures,” *Japanese Journal of Applied Physics*, vol. 44, pp. L1478–L1480, 2005.

[85] G. H. Jessen, R. C. Fitch, J. K. Gillespie, G. Via, A. Crespo, D. Langley, D. J. Denninghoff, M. Trejo, and E. R. Helle, "Short-channel effect limitations on high frequency operation of AlGaN-GaN HEMTs for T-gate devices," *IEEE Transaction on Electron Devices*, vol. 54, pp. 2589–2597, October 2007.

[86] Y. Awano, M. Kosugi, K. Kosemura, T. Mimura, and M. Abe, "Short-channel effects in subquarter-micrometer-gate HEMTs: simulation and experiment," *IEEE Transactions on Electron Devices*, vol. 36, no. 10, pp. 2260 –2266, October 1989.

[87] K. Nummila, A. Ketterson, S. Caracci, J. Kolodzey, and I. Adesida, "Short-channel effects in sub-100 nm GaAs MESFETs," *Electronics Letters*, vol. 27, no. 17, pp. 1519 –1521, August 1991.

[88] A. Endoh, Y. Yamashita, K. Hikosaka, T. Matsui, S. Hiyamizu, and T. Mimura, "Threshold voltage shifts in decananometre-gate AlGaN/GaN HEMTs," *Electronics Letters*, vol. 42, no. 8, pp. 490 – 492, April 2006.

[89] A. Kuliev, V. Kumar, R. Schwindt, D. Selvanathan, A. M. Dabiran, P. Chow, and I. Adesida, "0.15 μm gate-length AlGaN/GaN HEMTs with varying gate recess length," *Solid-State Electronics*, vol. 47, no. 1, pp. 117 – 122, 2003.

[90] M. Higashiwaki, N. Onojima, T. Matsui, and T. Mimura, "High f_t and f_{max} AlGaN/GaN HFETs achieved by using thin and high-Al-composition AlGaN barrier layers and Cat-CVD SiN passivation," *Physica status solidi (a)*, vol. 203, no. 7, 2006.

[91] J. Ayubi-Moak, D. Ferry, S. Goodnick, R. Akis, and M. Saraniti, "Simulation of ultrasubmicormeter-gate InAlAs-InGaAs-InAlAs-InP PHEMTs using a full-

band monte carlo simulator," *IEEE Transactions on Electron Devices*, vol. 54, no. 9, pp. 2327 –2338, September 2007.

[92] D. J. Denninghoff, S. Dasgupta, D. F. Brown, and U. K. M. S. Keller J. Speck, "Analysis of effect of source-drain spacing of N-polar GaN MIS-HEMTs with 310 Ghz fmax," *Proceedings of the 17th International Conference on Electron Dynamics in Semiconductors, Optoelectronics, and Nanostructure, Santa Barbara, CA, USA*, August 2011.

[93] Y. Ando, Y. Okamoto, H. Miyamoto, T. Nakayama, T. Inoue, and M. Kuzuhara, "10-W/mm AlGaNGaN HFET with a field modulating plate," *IEEE Electron Device Letters*, vol. 24, no. 5, pp. 289–291, May 2003.

[94] A. Chini, D. Buttari, R. Coffie, S. Heikman, S. Keller, and U. K. Mishra, "12 W/mm power density AlGaN-GaN HEMTs on sapphire substrate," *IEEE Electron Device Letters*, vol. 40, pp. 73–74, January 2004.

[95] Y.-F. Wu, D. Kapolnek, J. Ibbetson, P. Parikh, B. Keller, and U. Mishra, "Very-high power density algan/gan hemts," *IEEE Transactions on Electron Devices*, vol. 48, no. 3, pp. 586 –590, March 2001.

[96] R. Vetury, N. Q. Zhang, S. Keller, and U. K. Mishra, "The impact of surface states on the dc and rf characteristics of AlGaN-GaN HFETs," *IEEE Transactions on Electron Devices*, vol. 48, no. 3, pp. 560–566, March 2001.

[97] B. Green, K. Chu, E. Chumbes, J. Smart, J. Shealy, and L. Eastman, "The effect of surface passivation on the microwave characteristics of undoped AlGaN/GaN HEMTs," *IEEE Electron Device Letters*, vol. 21, no. 6, pp. 268–270, June 2000.

[98] M. Ha, S. Lee, S. Kim, Y. C.M., and M. Han, "Ni/au schottky gate oxidation and bcb passivation for high-breakdown-voltage algan/gan hemt," *Superlattices and Microstructures*, vol. 40, no. 4-6, pp. 562 – 566, 2006.

[99] H. Wang, J. W. Chung, X. Gao, S. Guo, and T. Palacios, "Al_2O_3 passivated InAlN/GaN HEMTs on SiC substrate with record current density and transconductance," *Physica Status Solidi (c)*, vol. 7, no. 10, pp. 2440–2444, 2010.

[100] J. Gillespie, R. Fitch, J. Sewell, R. Dettmer, G. Via, A. Crespo, T. Jenkins, B. Luo, R. Mehandru, J. Kim, F. Ren, B. Gila, A. Onstine, C. Abernathy, and S. Pearton, "Effects of Sc_2O_3 and MgO passivation layers on the output power of AlGaN/GaN HEMTs," *IEEE Electron Device Letters*, vol. 23, no. 9, pp. 505 – 507, September 2002.

[101] C. Liu, E. Chor, and L. Tan, "Enhanced device performance of algan/gan hemts using HfO_2 high-k dielectric for surface passivation and gate oxide," *Semicond. Sci. Technol.*, vol. 22, pp. 522 – 527, 2007.

[102] H. Sun, A. Alt, H. Benedickter, E. Feltin, J.-F. Carlin, M. Gonschorek, N. Grandjean, and C. Bolognesi, "205-GHz (Al,In)N/GaN HEMTs," *IEEE Electron Device Letters*, vol. 31, no. 9, pp. 957–959, September 2010.

[103] Y. Pei, S. Rajan, M. Higashiwaki, Z. Chen, S. DenBaars, and U. Mishra, "Effect of dielectric thickness on power performance of algan/gan hemts," *IEEE Electron Device Letters*, vol. 30, no. 4, pp. 313–315, April 2009.

[104] Z. H. Liu, G. I. Ng, S. Arulkumaran, Y. K. T. Maung, K. L. Teo, S. C. Foo, and V. Sahmuganathan, "Comprehensive study on the bias-dependent equivalent-circuit elements affected by PECVD SiN passivation in AlGaN/GaN HEMTs,"

IEEE Transactions on Electron Devices, vol. 58, no. 2, pp. 473–479, February 2011.

[105] F. A. Marino, D. A. Cullen, D. J. Smith, M. R. McCartney, and M. Saraniti, "Simulation of polarization charge on AlGaN/GaN high electron mobility transistors: Comparison to electron holography," *Journal of Applied Physics*, vol. 107, p. 054516, 2010.

[106] J. Chung, O. Saadat, J. Tirado, X. Gao, S. Guo, and T. Palacios, "Gate-recessed InAlN/GaN HEMTs on SiC substrate with Al_2O_3 passivation," *IEEE Electron Device Letters*, vol. 30, no. 9, pp. 904–906, September 2009.

[107] S. Taking, A. Banerjee, H. Zhou, X. Li, A. Khokhar, R. Oxland, I. McGregor, S. Bentley, F. Rahman, I. Thayne, A. Dabiran, A. Wowchak, B. Cui, and E. Wasige, "Surface passivation of AlN/GaN MOS-HEMTs using ultra-thin Al_2O_3 formed by thermal oxidation of evaporated aluminium," *Electronics Letters*, vol. 46, no. 4, February 2010.

[108] C. Choudhry and V. K. Arora, "Field dependace of drift velocity and electron population in satellite valleys in GaAs," *Electronics Letters*, vol. 22, no. 5, pp. 271–272, February 1986.

[109] P. G. Sverdrup, Y. S. Ju, and K. E. Goodson, "Sub-continuum simulations of heat conduction in silicon-on-insulator transistors," *Journal of Heat Transfer*, vol. 123, no. 1, pp. 130–137, 2001.

[110] E. Pop, R. Dutton, and K. Goodson, "Detailed heat generation simulations via the monte carlo method," *International Conference on Simulation of Semiconductor Processes and Device (SISPAD)*, September 2003.

[111] J. Rowlette, E. Pop, S. Sinha, M. Panzer, and K. Goodson, "Thermal simulation techniques for nanoscale transistors," in *ICCAD '05: Proceedings of the 2005 IEEE/ACM International conference on Computer-aided design*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 225–228.

[112] A. Matulionis, J. Liberis, I. Matulioniene, H.-Y. Cha, L. F. Eastman, and M. G. Spencer, "Hot-phonon temperature and lifetime in biased 4H-SiC," *Journal of Applied Physics*, vol. 96, no. 11, pp. 6439–6444, December 2004.

[113] L. Ardaravicius, A. Matulionis, J. Liberis, O. Kiprijanovic, M. Ramonas, L. F. Eastman, J. R. Shealy, and A. Vertiatchikh, "Electron drift velocity in AlGaN/GaN channel at high electric fields," *Applied Physics Letters*, vol. 83, no. 19, pp. 4038–4040, November 2003.

[114] M. Sadiku, *Numerical Techniques in Electromagnetics*, 2nd ed. Boca Raton, FL: CRC Press, 2001.

[115] J. Proakis and D. Manolakis, *Digital Signal Processing, 4th edition*. Upper Saddle River, NJ: Prentice Hall, 2007.

[116] M. S. Nakhla and J. Vlach, "A piecewise harmonic balance technique for determination of periodic response of nonlinear systems," *IEEE Transactions on Circuits and Systems*, vol. cas-23, no. 2, pp. 85–91, February 1976.

[117] R. Kamoua, "Monte Carlo-based Harmonic-Balance technique for the simulation of high-frequency TED oscillators," *IEEE Transactions on Microwave Theory and Techniques*, vol. 46, no. 10, pp. 1376–1381, October 1998.

[118] R. Lipsey, S. Jones, J. Jones, T. Crowe, L. Horvath, U. Bhapkar, and R. Mattauch, "Monte Carlo Harmonic-Balance and drift-diffusion harmonic-balance analyses of 100-600 GHz schottky barrier varactor frequency multipliers," *IEEE Transactions on Electron Devices*, vol. 44, no. 11, pp. 1843–1850, November 1997.

[119] F. Bertazzi, F. Bonani, S. D. Guerrieri, and G. Ghione, "Large-signal device simulation in time- and frequency-domain: a comparison," *The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 27, no. 6, pp. 1319 – 1325, June 2008.

[120] H. I. Fujishiro, S. Narita, and Y. Tomita, "Large signal analysis of AlGaN/GaN-HEMT amplifier by coupled physical device-circuit simulation," *Physica Status Solidi (a)*, vol. 203, no. 7, pp. 1866 – 1871, 2006.

[121] W. Contrata, Y. Ando, Y. Hori, and M. Kuzuhara, "Intermodulation distortion simulation using physical GaAs FET model," in *Microwave Symposium Digest*, vol. 2. IEEE, 2001, pp. 819 – 822.